DOCUMENT RESUME

ED 303 486 TM 012 680

AUTHOR Arter, Judith A.
TITLE Curriculum-Referenced Test Development Workshop Series: Workshops One through Three.
INSTITUTION Northwest Regional Educational Lab., Portland, OR. Assessment and Evaluation Program.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 88
CONTRACT 400-86-0006
NOTE 280p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC12 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Inservice Teacher Education; Item Banks; Multiple Choice Tests; School Districts; Teacher Made Tests; *Teacher Workshops; *Test Construction
IDENTIFIERS Curriculum Based Assessment; *Curriculum Referenced Tests; *Test Specifications

ABSTRACT

This set of materials represents the first three workshops in a series of five designed to assist school districts and educators to develop their own curriculum-referenced tests. The series has been assembled to provide school districts with a relatively inexpensive test development method. The series is designed to reduce costs by pooling resources in terms of training and instrument development. Although many of the concepts presented can be used for informal, daily classroom assessment, the main focus is the development of more formal assessment systems (unit, year, or course-end tests) and/or diagnostic systems. The three workshops include an introductory session, a session on developing test specifications, and a item pool development session. The two workshops not outlined cover pilot testing and finalizing assessment materials. The entire series is designed to occur over a period of 1 year or more. Each workshop presents information and practice on one step of the test development process. Participants then finish that step before they proceed to the next workshop. At the end of the series, participants have one or more instruments pilot-tested and ready to use. Although many of the instruments are multiple-choice, other formats are encouraged as appropriate. Presenters' outlines, participant handouts, and copies of transparencies are provided for each of the three workshops. Samples of letters to participants are appended. (TJH)

CURRICULUM-REFERENCED TEST DEVELOPMENT WORKSHOP SERIES

WORKSHOPS ONE THROUGH THREE

Developed By:

Judith A. Arter

November 1988

Sponsored by OERI Office of Educational Research and Improvement U.S. Department of Education

Northwest Regional Educational Laboratory
101 S.W. Main Street, Suite 500
Portland, Oregon 97204

Test Center

November 1988

This publication is based on work sponsored wholly or in part by the Office of Educational Research and Improvement (CERI), U.S. Department of Education under contract Number 400-86-0006. The content of this publication does not necessarily reflect the views of the department or any agency of the U.S. government.

CURRICULUM-REFERENCED TEST DEVELOPMENT WORKSHOP SERIES

WORKSHOPS ONE THROUGH THREE

Developed By:

Judith A. Arter

November 1988

Test Center
Evaluation and Assessment Program
Northwest Regional Educational Laboratory
101 S.W. Main, Suite 500

TABLE OF CONTENTS

INTRODUCTION

This set of materials represents the first three workshops in a series of five designed to assist districts and other educators to develop their own curriculum-referenced tests (CRTs). The series has been put together because of a felt need for districts to have a less expensive test development alternative. The attempt is to reduce costs by pooling resources in terms of training and developing instruments.

Although many of the concepts presented can be used for informal, daily classroom assessment, the main focus is the development of more formal assessment systems -- unit, year or course-end tests; or, perhaps, diagnostic systems.

The entire workshop series is designed to occur over the period of a year or more. Each workshop presents information and practice on one step of the test development process. Participants then finish that step before the next workshop. At the end of the series participants have one or more instruments pilot-tested and ready to use. Although many of the instruments are multiple-choice, we encourage other formats (for example, essay, performance, teacher-ratings) when they are appropriate to adequately measure the outcome intended.

The content of each workshop in the series is as follows:

Workshop #1 -- Introduction To The Workshop Series

This session describes the test development process and related issues so that participants can decide whether the series is something that is appropriate for them at the current time. The session also emphasizes the nature of cooperative efforts.

Workshop #2 -- Developing Test Specifications

This session covers purposes for testing, reviewing the curriculum for test development, and developing a blueprint for a test. Participants will begin to develop their testing plans and blueprints. This process is completed as homework.

Workshop #3 -- Developing Item Pools

This session assists participants in assembling pools of test questions that match their testing blueprint. Help is given on finding items, writing items, formatting items, obtaining input on items and the logistics of handling large numbers of items. We encourage participants to use existing item pools as much as possible. The item pools are completed as homework.

Workshop #4 -- Pilot Testing

Participants develop plans for pilot testing their instruments -- who to assess, how to put together the pilot test versions and what other information to collect. Help is also given on what statistics to generate on the tests. The pilot testing is completed as homework.

Workshop #5 -- Finalizing Assessment Materials

Participants learn how to use various types of statistics to refine their tests. Assistance is given on how to finalize the instruments, what to put in a technical report, what to put in a user's manual, what to put in a district testing policy and how to report and use results effectively.

Although the above represents the default content of each workshop in the series, the participants should have a hand in moulding the series to fit their interests and needs. Every series we have done so far has been different in terms of added or rearranged content. Sometimes, for example, participants want to add an extra session on the change process so that they can build commitment in their local situations. Other times everyone wants to work on a common subject area, so workshop content is much more specific.

Presenters need to have the following knowledge and resources:

1. How to build assessment instruments of various types in both structured and open-ended formats.

2. How assessment is used in effective schools and how change occurs in schools.

3. Software available for test development.

4. Where to find pools of items in various subject areas. (It helps to have access to these pools.)

5. Facilities to run statistical analyses on information generated from the assessment process.

This document contains the workshop materials for the first three workshops in the series. For each workshop we have provided presenter's outlines, participant handouts and hand copies of transparencies. Sample letters written to participants between sessions are provided at the end.

WORKSHOP #1

INTRODUCTION TO THE WORKSHOP SERIES

PRESENTOR'S OUTLINES

Introductory Workshop -- Overview of the Day
Presenter's Outline

Purposes: To describe the purposes for this workshop and the entire series.
To provide an overview of the current day's activities
To introduce the presenters, state department personnel and participants.

General Information on Workshop #1:

Supplies needed for the day are nametags and colored dots (for small group dicussions); blank paper and pencils; folders or notebooks in which participants put materials; and chart packs and pens (one for each 5-7 participants).

Presenters for this workshop should have background and experience in two areas: developing tests (particularly curriculum-aligned tests with districts) and experience in the change process in schools (especially practical experience in using test scores positively).

It is a good idea to involve local or state representatives as cospnsors and/or copresenters. Leave time for them to do introductions, announcements, how this activitiy fits in with local priorities, etc.

Format: Lecture/discussion

Materials: Workshop Title transparency
Agenda handout and transparency
Overhead and projector

Presentation Time: 20 minutes

Points to Make:

1. Introduce presenters, state department personnel, etc. This includes time for orienting comments by cosponsors.

2. The workshop series concept came to be because of a felt need that small and medium (indeed all) districts want to develop tests that are aligned to their own curricula, but since test development is expensive and requires some expertise, many districts cannot do it. The Evaluation and Assessment Program at the Lab is exploring ways to decrease the cost to districts for such test development. Briefly describe the series and how it can be tailored for individual needs. The end product of the series is usually one or more assessment instruments that has been pilot tested and is ready for use. Emphasize the various ways that participants can work together -- share items, work in the same subject area, cooperative pilot testing, etc. Describe previous series and their results.

   The purpose of the first workshop is to discuss the test development process and related topics so that participants can decide if this is something they want to do at the current time.

3 The day has three basic aspects:

   a. Information on and overview of the technical aspects of test development -- how to write items, item statistics, pilot testing, reporting, norms, setting passing

scores, etc. This is done so that participants know what they need to do to develop a good test.

b. Information on and discussion of the social and political aspects of testing, test development and test results use. This will entail discussions of the change process, how assessment information fits in with effective schooling practices, and purposes for testing. This is a major cluster of topics not only because accessible information figures so prominently in the literature on effective schooling, but also because test development and use represents a change it and of itself. In order to be successful with this effort, we need to attend to how to implement change.

   Management Driven Instruction (MCI) is controversial right now (e.g., Popham v. Sheppard at AERA in April, 1988). There is lots of potential but there are also pitfalls. The most intractible pitfalls seem to be social and political rather than technical. There is a tension between testing as a science and testing to serve political ends. Examples of social and political concerns are perception of use of results, and perception of the usefulness of results.

c. District discussion and sharing. This facilitates networking, provides the opportunity to discuss individual needs and concerns and help each other solve problems, and helps provide information to promote tailoring the workshop series.

4. In the morning we will discuss the vision -- why having assessment information is important, how do we plan for this change, etc. In the afternoon we will plan -- where do we go from here.

Definitions of Terms
Presenter's Outline

Purposes: To provide common definitions for testing terms to be used throughout the workshop series.
To provide information to participants on types of tests and distinctions between tests.
This builds into later topics.

Format: Lecture/discussion

Materials: Definitions handout and transparency
Overhead and projector

Presentation Time: 20 minutes

Content of Presentation:

1. **Standardized** means that the testing and scoring processes are uniform. One gives the tests in the same way to all examinees and interprets the results in a uniform manner. A standardized test does not necessarily mean a published norm-referenced test (NRT), although the term is sometimes used that way. The curriculum-referenced tests (CRTs) participants develop will probably be standardized.

2. A **survey test** is one in which many objectives are tested. Usually, in order to keep test length down, this implies only a few questions per objective (1-3). Such tests are used for general assessment of student learning over broad areas of content. Because of the few number of questions per objective, results cannot be interpreted at the objectives level, but only at the skill cluster level.

3. At the other end of the content coverage continuum are tests in which a few objectives are tested in greater detail. Usually there are 6-15 questions per objective. Unit tests, mastery tests and diagnostic tests are examples of this. Results can be interpreted at the objectives level

4. A **norm-referenced test** is one in which scores are available that compare student performance to that of other similar students. Local or "national" norms can be developed for most tests. The term is sometimes used to refer to standardized, published tests, but really, most tests can be normed. The term is also sometimes used to refer to the method by which the test questions are selected for use on the test -- questions are chosen to "spread" out student scores so that student standing in the group can be determined.

5. A **criterion-referenced test** is one for which scores are available that compare student performance to a criterion for mastery. Standards for mastery can be developed for any test. The term is sometimes applied to tests which were developed to show instructional progress rather than to compare students to each other. Thus, test questions would be selected for the test to measure important learning outcomes, regardless of whether or not they spread students out in terms of achievement. The term is also sometimes applied to a test development process in which the learning domain to be measured is very carefully defined so that the sample of questions on the test accurately represents the domain.

6. A **curriculum-referenced test** is one which has been developed specifically to match up with a particular set of curriculum objectives. As such it can be a survey test or a diagnostic test, and it can have norms or have set criteria for mastery. The term mainly describes the degree of match between the test content and a particular curriculum. **We will use the term CRT to refer to curriculum-referenced tests not criterion-referenced tests.**

Other points to make:

1. Different test types (e.g., nationally developed survey tests and locally developed end of course/year tests) serve different purposes. (For example, national survey tests are good for seeing how one is doing compared to others, diagnosis on a broad scale, selecting students for special programs, and checking on how the local curriculum is faring in developing the types of skills deemed important across districts and textbook publishers. Local CRTs are good for individual and group instructional planning on local objectives, seeing the extent to which students are learning what "we" want them to, selecting students for programs and certification.) One is only better than another depending on what information is needed. There is nothing magical about CRTs.

2. This implies that users know what purposes for test information they have so that they can rationally decide on the best test mix to serve all their information needs. (The next section more fully discusses purposes for CRTs.)

3. One test cannot serve all testing purposes because the design of the test will vary depending on the intended use. (For example, survey tests are broad in scope and shallow in depth; diagnostic tests are narrow in scope and deep.)

4. A good testing system would probably be a combination of a NRT and a CRT in order to answer the majority of questions that local districts ask.

5. Currently, districts usually end up with survey tests when they develop their own tests. This is because of constraints on time and money.

Why Have Curriculum Aligned Tests?
Presenter's Outline

Purposes: To discuss the reasons for assessment information in general and curriculum-aligned assessment information specifically.

Format: Lecture/discussion

Materials: Beverly Anderson's article on purposes for testing and the change process.
NWEA's handout that lists the aspects of the effective schooling literature that involve having information.
Overhead projector and screen

Presentation Time: 30 minutes

Content of Presentation: (This information is partially based on prior presentations by Allan Olson of Hillsboro School District in Oregon, Dick Sagor of West Linn School District in Oregon and Wayne Neuberger of the Oregon State Department of Education.) Points to cover include:

Reasons for wanting curriculum-referenced tests:

1. Assessment information fits into several places in the research on effective schools. (See the handout entitled *Effective Schooling Practices Related to Assessment.*) This handout points out that assessment information is essential for monitoring progress so that instruction can be adjusted and students can move on when they're ready; grouping students to fit instructional needs; monitoring the status of the educational program as a whole to see the extent to which the school is accomplishing what it wants to accomplish, spotting potential problem areas, and adjusting the curriculum; and targeting program improvements.

2. Information is used for diagnosis and instructional planning. This can create power for teachers and principals. But these individuals need to buy into the notion of this potential power. This will not happen if they perceive the testing system to be serving mostly others and not them. CRT use should focus on looking at instruction and the curriculum and getting the information to those that need it.

3. Although teacher and administrator intuition about progress can be useful, intuitions can be wrong, and what happens when one set of intuitions is overruled by another set? When one has data to support and form a position, one has a stronger case.

4. Information can be motivating. Motivators for professionals include a sense of achievement; recognition; a feeling of being in control; a feeling of competence; and a feeling of usefulness. Information can serve these functions.

Purposes served by CRTs:

1. Bev Anderson's article does a nice job of summarizing the various uses for test scores in general. CRTs are most useful for local instructional planning, curriculum revision and seeing the extent to which local goals are being achieved. NRTs serve to provide a view of how achievement locally compares to that

nationally and provides an independent look at whether the local curriculum should be revised. (Reasons for this include extent and nature of match to local curriculum and frequency of testing. The point could be made that it is not necessarily true that NRTs do not match the local curriculum. The nature of the match, however, is generally more global. Need for information on specific skills may be lacking.) A good testing system probably includes both CRTs and NRTs.

2. The handout entitled *Effective Schooling Practices Related to Assessment* also does a nice job of showing how a CRT system should be designed. This also demonstrates again how the perception of a test as being "high stakes" can ruin its use for instructional purposes.

Where Have We Been, Where Are We Going?
Small Group Discussion
Presenter's Outline

Purposes: To provide an opportunity for participants to share their test development experiences and get to know each other. Since we are trying to promote cooperative efforts we are trying to set the tone early. This will also serve a networking function, and will help the presenters to begin seeing how the workshop series should be tailored to the needs of the current participants. Finally, districts can learn a great deal from each other and rarely have the opportunity to talk to each other.

Format: Small and large group discussion

Materials: Small Group Discussion handout
Testing Information Sheet handout
Chartpack and pens for each group of 5-6 participants
Pencils for participants

Time: 60 minutes

Sequence of Events:

1. Use about 10 minutes to fill out the Testing Information Sheet. Representatives of the same district should fill this out jointly before breaking into small groups. The Testing Information Sheets will be compiled and mailed out to all participants for networking purposes; they are also used to help tailor the workshop series.

2. Break the group up into small groups of about 5-6 persons each. Representatives from the same district should split up. Colored dots on nametags facilitate this process.

3. Use about 30 minutes (5 minutes per person) to discuss local test development efforts in terms of purposes for testing, development procedure, use, development/ongoing cost, benefits/problems, political issues, future plans, etc.

   One person should be designated a spokesperson to report back to the whole group general impressions of the discussion and sources of resources for test development. General impressions could include such things as striking similarities and differences in purposes, what the tests look like, political implications, technical issues, money, etc. The spokesperson does not need to record specific activities from each district because this information will be collected when individuals fill out the Testing Information Sheet.

4. Reconvene the large group. Get a show of hands corresponding to the areas on the Testing Information Sheet. (The purpose of this is so that everyone knows who has done what in the past and knows about current efforts similar to their own.)

5. Finally, have one representative from each group share their general impressions of the discussion at their table. This should take about 10 minutes.

Typical Responses:

Typically, about 1/3 to 1/2 of the participants have already initiated plans for test development. Of these, a few have developed tests before. However, most of them are at the curriculum development stage and are looking for help with the next logical step of test development. Participants at this stage have experienced or are experiencing attempts to build local commitment, develop resources and sell teachers and the community. If participants are at this stage and do not report issues surrounding building local commitment, they probably do not have in place the structure needed to make the system successful ultimately. It is good for these participants to have the opportunity to discuss issues that have arisen and engage in group problem-solving. Issues that arise are fears about how the results will be used, how to sell others on the idea, why would districts want to have locally aligned tests, how to convince people of the need and where to get resources. The next section on the change process discusses some of these points. The previous section on why CRTs? discussed other of these points.

Most of these participants will form district teams and proceed through the workshop series ending up with district tests at the end.

The remainder of the participants want their districts to move in the direction of CRTs but are still in the throws of trying to make decisions and set directions. The discussions on how others have moved through this stage are useful for this group. A portion of this group will continue on through the workshop series just to learn how to develop tests even though they will not end up with district tests at the end.

In terms of subject areas and grade levels participants are typically all over the place -- from science to reading and from grade K to 12.

It is also possible to reconfigure the workshop series if enough districts are at the point of indecision. An extra workshop could focus on implementing change.

The Test Development Process Part 1 --
The "People" Aspect
Presenter's Outline

Purposes: To emphasize that test development and the use of assessment results constitutes a "change;" Such a change has to be handled like other system changes.
To describe what some of the people concerns are and how they can be dealt with.

Format: Lecture/discussion

Materials: Criteria For Analyzing Plans To Develop CRTs handout and transparency
Management Support transparency
Overhead projector and screen

Presentation Time: 60 minutes

Content of Presentation: (This information is based on presentations by Dick Sagor of West Linn school district in Oregon, and Allan Olson of Hillsboro school district in Oregon.)

Developing tests and promoting information based decision making is a change undertaken to improve instruction and have a positive effect on students. Success depends on people -- doing the development work, using the results in the way intended, etc. The success of the project, having CRT test results used successfully to inform instruction, depends on several key ingredients. These have been generated from two decades of examining what is necessary in order to have change succeed. (The handout was based on one developed by NWEA's Science Project.)

1. Does management understand and support the improvement effort?

   There are three stages in the change process:

   a. **Initiation.** Historically this is where administrators are most involved. They are the people with great ideas. They go a conference, hear about something terrific, come back to a faculty meeting, and sell everybody on how neat this idea is. (E.g., team teaching, peer coaching, assertive discipline, etc.) However, what can happen is that once everyone is convinced the leader checks out or goes to another conference and comes back with another great idea. About 90% of projects die out at this point unless leadership continues.

   b. **Implementation.** This is the hard part. A project will only succeed if there is leadership during the middle years.

   c. **Incorporation.** This is where the new idea becomes the way we do it here. At this point the need for leadership levels off.

   There are two types of leadership that are needed during this process and the relative need for them varies by the stage of implementation:

1. **Directive support.** This is most important at the implementation stage. You need a knowledgeable person there directing the work. You need experts to get it to go. This need drops off over time.

2. **Emotional support.** People need pats on the back even after the need for directive support lessens.

The implications of this paradigm are:

- o A commitment to an improvement is a long-term matter on the part of the administration.
- o People need to know that this is where attention will be focused for awhile.
- o Therefore, do not try to do too many things at once. Do one or two and try to do them well.
- o If the program dies because of lack of support and direction from the administration, it will reinforce the view of colleagues not to take new programs seriously and not to invest too much time in them.
- o Adequate support for the improvement involves participation of those affected, time to do tasks well, adequate training, and time to practice. These are all concrete evidences that there is commitment to the project. (This also covers points #3 and 7 on the handout.)
- o There should be incentives to participate and ongoing acknowledgements of accomplishments. (This also addresses #5 on the handout.)
- o The leader has to pay attention to incidents (individual interactions and happenings). The perception of the project by teachers and others will depend on incidents. For example, a negative impression occurs if tests are scored incorrectly. Other incidents are teacher conversations, noticing that a task becomes easier, a note in a mailbox, etc. It is the little day-to-day things that count.

2. Is there a felt need by the staff? Teachers do not list their #1 need as being a CRT system. We need to help them recognize what they want -- have increased power to help students do better. A good CRT system can do this. Early concerns by teachers tend to be informational and personal -- what is the change and how will it affect me? (Task and impact concerns come later.) Some issues, concerns and solutions are:

   a. Ways to build felt need are to do a formal needs assessment; have a team member by the mailboxes in the morning to have discussions with colleagues to see what their concerns are; get people together to look at data (have them interpret it themselves and discuss implications); quickly find an application of CRT results that saves people time (e.g., better placement of students into courses or instructional materials); get users involved in the planning and construction of the tests and reports.

b. How will the test results be used? We advocate use of CRT results for instructional planning and curriculum review at the classroom, building and district level. In order for this to work the teachers must perceive the tests to be for them, and not to be checks on them. There is ample evidence that the perception of "high stakes" leads to pressure which leads to perversion of the testing and information flow process. This can result in information that is not useful and also not used. This is perhaps the biggest issue in testing today -- how to use assessment results in supportive and not punishing ways.

Trust has to built and sometimes it takes years. It is helpful to have a stated policy as to the intended uses of the test results. Reporting and interpretation of results needs to support this trust. The teachers also need to see that teaching to the curriculum is teaching to the test, because the test mirrors the curriculum.

It is not fair to directly compare teachers on test scores because of the other variables that can cause achievement levels that have nothing to do with instruction, such as parental support, SES level of students, district resources, etc. Therefore, it may be fairer to look at pre-post gain and longitudinal results rather than point in time comparisons.

c. Will having tests that everyone uses result in restricting the curriculum? This can, in fact, happen, especially if the test is perceived as "high stakes." Some solutions are to include a broad range of instances of skill application on a test, rotate test questions, use other assessment procedures than multiple-choice, and rotate the skills assessed. There is also some evidence that skills are only added and not taken away. That is, teachers will not tend to replace instruction on certain skills with those on the test, but will rather add any interesting skills from the test to those already being addressed.

d. Will the curriculum be trivialized because it is difficult to assess many important educational outcomes? Content that is harder to measure includes higher-order thinking skills, citizenship and affective areas such as self-concept and attitude toward learning. Trivialization can, in fact, happen, especially if the test is seen as being "high stakes." Test content indicates the educational priorities of the district. If the test is going to impact instruction, we want the right priorities emphasized. Some solutions are to rotate content, and use other assessment formats than multiple-choice.

e. Will there be less freedom for teachers? Teachers may feel that they need to get from point A to point B in a certain amount of time and therefore do not have time for teachable moments or emphasizing other outcomes. One solution is not to emphasize facts on the test but emphasize concepts, perspectives and the ability to use and analyze information. Also the test should be a sample from the entire domain of skills that are **locally identified to be important**. Thus, good instruction targeted at the domain should be reflected in the sample of applications on the test.

f. Why should I invest time on this innovation because most innovations pass by quickly? This concern can be addressed by the comments under leadership above.

g. Where do we get the time to do this? This has to be a top priority commitment on the part of administration if they are committed to the success of the project.

h. There is already too much testing; we don't need any more. One solution is to survey the buildings to determine all the tests that are being used and their purposes. Perhaps this could be streamlined.

Concerns such as these need to be known and targeted. Activities should be planned to legitimize and deal with these concerns.

4. Does the improvement fit local school conditions? Some aspects of the plan should be left up to individual buildings and classrooms. For example, if tests are going to be given in all high school courses, have each department decide what the tests should look like, how retesting will occur, what the passing score(s) should be and how results will be reported.

5. Are there incentives to participate? Teachers will take on more if they are convinced that in the long run it will save time and make instruction better. Teachers are already overworked. What are the carrots?

6. Is staff development adequate? This means training and supervised practice.

8. It is important to see some results quickly. These should be things that make people's jobs easier or better. For example, saving time assigning students to classes, providing information that is useful to teachers, etc.

9. Are all plan elements ready? Are the tasks complete, clear and well sequenced? Is there adequate time to do each step? Are tasks sequenced to keep the project fresh? Are responsibilities assigned? Are there adequate resources? Is the vision of what the CRT system should look like and what it will accomplish clear? Is policy being planned (district approach; uses of results; subjects; grades).

The Test Development Process Part 2 --
Test Development Steps
Presenter's Outline

Purposes: To let participants know the tasks involved in a test-development project; what some of the pitfalls are and the resources/personnel needed.

Format: Lecture/discussion

Materials: Test Development Steps handout and transparency
Overhead projector and screen

Presentation Time: 75 minutes

Content of Presentation:

Step 1: **Decide on the purposes for testing**

Purposes for testing will determine what the tests will look like, who the results will be reported to, how the results will be used, and the resources needed. The handout "Purposes for Testing" illustrates how differences in purposes influence the way the test is constructed. Everyone needs to agree on these purposes ahead of time as part of the process of building commitment and ownership. This process is heavily "people-oriented." Deciding on the purposes for testing will take more or less time depending on where the district is in the process.

We recommend against using the test for teacher evaluation or any other purpose seen as "high stakes." It is difficult to have the same test serve both as an impartial information sources for teachers and to evaluate teachers. The teachers need to own the test in order to value the results.

Step 2: **Develop Testing Blueprint**

The purpose of the blueprint is to specify what the test and test questions will look like. This document guides the development of item pools. The blueprint specifies total test length; the objectives to be covered; the number of questions to be developed on each objective; how objectives relate to each other; and how each objective will be assessed. Ideas for items and specific content is often also included.

Issues are breadth v. depth; specificity of curriculum objectives; how objectives relate to each other; how to decide which objective to assess; how to assess things not easily assessed in objective-format; and reductionism.

The test specifications should be reviewed by all interested parties to build ownership and commitment -- teachers, administration, students? With curriculum objectives in pretty good shape, it should take the committee three to four 2-hour meetings (after training) to development specifications and reviewers about 1/2 day per test to provide input. Add to this the time and cost of word processing, duplication and editing.

Step 3: **Develop item pools.**

An initial pool of items, prompts, scenarios, etc. is developed for each objective. About 1 1/2 to 2 times the number of items needed is generated. Use preexisting items when possible -- teachers, item banks. Only write when you have to. These items should be reviewed for match to objectives, bias, and technical adequacy. Issues are adequate sampling of domains to be assessed; and approaches that illuminate instruction.

Item pools are developed by relevant groups of teachers and reviewed by all interested teachers. In the basic skills, for a 200-item pool, it should take two teachers about two days of pulling items and two days of writing items to develop a pool for review. Reviewing a 200-item pool carefully should take a teacher about 1/2 day. Add to this the cost of coding items, entering items, and editing items. These times will increase with other subject areas.

Step 4: **Developing pilot test forms**

Items marked by reviewers as being of highest priority are pilot-tested. You should pilot test about 125% to 150% the number of items that will be on the final form. The higher number applies to hard to define objectives. At least 50 students per course/grade should take the pilot form. More students are needed depending on other analyses. Pilot testing also requires administration instructions, sample items and graphics.

Teachers should review the pilot forms. Assembly of each pilot form should take 1 day of professional time and 2 to 3 days of support staff time. Administration time varies. Scoring time depends on the type of questions.

Dpending on the proposed use for the results, other special studies will need to be undertaken. For example, HOTS might require having students take the items orally.

Step 5: **Analyzing the results and refining the tests**

There are standard statistics for test items. These, plus teacher comments and observations, can be used to finalize the test items. Basic analysis is very quick with a canned program. Interpretation and deciding which items to keep and which to revise can take 1 to 3 days of professional time per test.

Depending on the special studies and the nature of the use of the test scores, more analysis might need to be done. For example -- test equating, standard setting, bias, relationship to other variables, and factor analysis.

Step 6: **Developing final forms and other documents**

Along with the final forms of the tests, you may want to develop technical reports, user's manuals, a district testing policy that outlines uses, purposes and philosophy, standard reporting formats and procedures, and standard procedures for reviewing and interpreting the results. You may also want to start an item bank. Sometimes you will pilot test a second time. Remember that test development is a process not an outcome. Tests are continually refined and revised.

What Do We Need To Get Where We're Going?
Small Group Discussion
Presenter's Outline

Purposes: To have district groups have planning time to think through what it would take for them to get their effort off the ground, and/or keep the effort moving long; and to discuss the advantages and possibilities of a cooperative effort.
Provide information to each other on concerns and issues.
Provide information to help tailor subsequent workshops in the series.
Model a process that participants could use in their own districts to help build commitment and plan.

Format: Small and large group discussion

Materials: Needs Brainstorming handout
Chartpack and pens for each district group
Pencils for participants

Presentation Time: 60 minutes

General Approach

When people want to undertake a large project, especially one for which commitment needs to be built, there are often obstacles. So we want to approach planning the task more in the spirit of "Where would breakthroughs need to occur for us to be able to get this project done?" Rather than "What obstacles do we need to overcome?" Sure, we know that given all we know now it isn't possible to get these tests developed in a year. But would you be willing to try to find creative solutions that would enable us to get the project done? Once people agree that it would be a good idea to have these tests, and that there is a committed group, then they have to outline where such breakthroughs would need to occur. These are the target areas for work and creative problem solving.

Sequence of Events:

1. Districts will stay in their district groups. Each group should have a chartpack and pen.

2. Districts will brainstorm what it would take, where breakthroughs would need to occur, to get their development effort off the ground; and/or keep it moving along. All responses are acceptable, regardless of level of detail, perceived triviality, perceived silliness, perceived redundancy, or perceived lack of central importance. Needs and concerns could include (but are not necessarily limited to) such things as: resources, time, information/expertise, equipment, political climate, local commitment, and need for local decisions (e.g., who, what, when, where and why). Areas could include teacher fears, how information will be used, cost, etc. It is also acceptable to be at the stage where you simply do not know whether this is something you want to do and want more information in order to decide.

3. Prioritize the areas in terms of which are most crucial to address in order to move ahead. List the top 10 concerns and needs. Appoint a spokesperson to report back to the large group.

4. Reconvene back into the large group. Have each group report back. See what commonalities and differences there are. Help each other problem solve.

Typical Responses:

Responses typically fall into three categories: building the vision, selling the vision, and implementing the vision. All these areas include finding the time to work on the project. Specifics have included:

Building the Vision

- o Purposes for testing, especially ways CRTs can be used
- o How to decide on uses and purposes
- o How use translates into how the test will look
- o Reporting to support uses
- o What a district policy statement on testing should look like
- o What a district handbook on testing should look like
- o How to cut down on testing
- o Time

Selling the Vision

- o More information on the change process, how to build felt need, how to sell the concept, how to respond to various concerns about testing, how to get started, who to talk to, etc.
- o How to develop resources
- o Time

Implementing the Vision

- o Expertise on test development
- o Where to get test items
- o Computer software to support test development
- o Help with curriculum development
- o How to write test questions
- o Time

Note: The first workshop series given did not include as much of the vision building and change process information as is included in the current materials. Because many of the responses to the Needs activity fell into Building and Selling the Vision categories, these sections were added to the introductory workshop. Consequently, the Needs activity in subsequent series reflected more interest in Implementing the Vision. However, the series presenters should be prepared to add an extra workshop on vision building and the change process if the somewhat brief treatment in Workshop #1 is not enough. The workshop series is designed to handle most of these topics. In addition, discussion of issues and concerns should be a formal part of each workshop in the series.

Proposed Workshop Series
Presenter's Outline

Purposes: To provide a default option for the workshop series.
To provide a tentative schedule to help participants decide whether this is something they want to do.
To provide a stimulus that participants can react to in light of their own discussion of needs.
To alter the content of the next workshop as needed.

Format: Lecture/discussion

Materials: Proposed Workshop Series handout and transparency
Session #1 Homework handout and transparency
Overhead and screen

Presentation Time: 30 minutes

Content of Presentation:

Before describing the proposed workshop series and homework mention the following caveat: The content of all workshops and the homework assignment today can be modified according to participant needs.

Each workshop will have planned time for participants to discuss issues and concerns that have arisen.

The initial plan for the series is:

1. Homework from the first session (see handout) would include:

   - Decide if you want to participant in the workshop series. Participation implies sending a team of 3-5 persons to each session; doing the homework in between sessions; providing the resources needed to complete the project (teacher release time; computer time; etc.) Consider the payoffs as well as amount of time, resources, and commitment needed to develop your own tests. Let _____ know by _____ if this is something you want to do

   - If you decide to continue, establish a team to guide the test development process. This team should be able to make decisions about the purpose of the test, subject areas to be assessed, and grades or courses to be included. A good team composition would be 3 to 8 central office personnel, principals and teachers.

   - Determine the purpose for your tests. This decision should involve policymakers as well as users of the tests. If different groups have different uses in mind for the tests, potential conflicts will occur down the road.

   - Decide on subject matter and grade levels to be included in testing. Consider the condition of the curriculum goals and objectives as well as readiness of the staff in making your selection. The curriculum goals or

objectives should be well understood and there should be common agreement as to their importance. In addition, there should be some confidence that the curriculum being taught corresponds relatively well with the written curriculum. If major curriculum work needs to occur, then testing may be premature at this time.

- o Establish a teacher review team to help with the test development process. The teachers should be well versed in the curriculum and able to work well with other staff members. The size of the team depends on the scope of the project.

- o Let _____ know by _____ purposes, subject area and grades. Also provide the list of names of persons who will attend the workshops. These persons should be a subset of the development team.

- o Be ready to bring to the next session the curriculum goals and objectives in the area to be assessed.

2. **Workshop #2: Developing Test Specifications** (1 to 2 days)

This session covers purposes for testing, reviewing the curriculum for test development, and developing a blueprint for a test. Participants will begin to develop their testing plans and blueprints. This process will be completed as homework.

3. **Workshop #3: Developing Item Pools** (1 to 2 days)

This session assists participants in assembling pools of test questions that match their testing blueprint. NWREL will assist districts in finding item banks to minimize the number of questions that need to be written. Help will be given on writing questions, formatting questions, obtaining input on questions and the logistics of handling large numbers of test questions. Development and review of item pools will be completed as homework.

4. **Workshop #4: Pilot Testing** (1 to 2 days)

Participants will develop plans for pilot testing their test questions -- who to test, developing test administration instructions, how to put together the pilot test forms and what other information to collect. Assistance will also be provided on how to score the results and what statistics to generate. Assembling pilot test forms and pilot testing will be completed as homework.

5. **Workshop #5: Finalizing Assessment Materials** (1 to 2 days)

Participants will bring their pilot testing results and will learn how to use item statistics to guide item revisions. Assistance will be provided on how to finalize the tests and other testing materials, what to put in a technical report, what to put in a user's manual, what to put in a district testing policy and how to report and use results effectively.

The task for the remainder of the time is to get a straw count on number of districts interested in continuing, tentative dates for sessions and how the content of at least the next workshop might be tailored.

PARTICIPANT HANDOUTS

CURRICULUM-REFERENCED TESTING SERIES

INTRODUCTORY WORKSHOP

(Date)
(Location)

Sponsored By:

_____ State Office of Public Instruction
(Other sponsors)

For more information call: _____

CURRICULUM-REFERENCED TESTING SERIES

INTRODUCTORY WORKSHOP

AGENDA

8:00 - 9:00	Registration and coffee
9:00 - 9:30	Introductions Overview of the day and purpose for each activity Definitions of testing terms
9:30 - 10:00	Why might districts want curriculum-aligned tests (CRTs)? What purposes do they serve?
10:00 - 11:00	Small and large group discussion and sharing on past, present and future district CRT plans and activities -- CRT experiences. (Plus break)
11:00 - 12:00	The test development process Part 1 -- The "people" aspect of successful and useful assessment and implementing change
12:00 - 1:00	Catered lunch
1:00 - 2:15	The test development process Part 2 -- The steps in developing a test
2:15 - 2:30	Break
2:30 - 3:30	Small and large group discussion -- Needs brainstorming. What do participants need in order to get their CRTs off the ground or keep their project moving along?
3:30 - 4:00	Planning the workshop series -- Proposed workshop sequence (and possible modifications based on expressed needs); what commitments districts need to make to continue with the series; tentative date for next workshop; homework.

Definitions

Standardized means that the testing and scoring processes are uniform. One gives the tests in the same way to all examinees and interprets the results in a uniform manner. A standardized test does not necessarily mean a published norm-referenced test (NRT), although the term is sometimes used that way. The curriculum-referenced tests (CRTs) participants develop will probably be standardized.

A **survey test** is one in which many objectives are tested. Usually, in order to keep test length down, this implies only a few questions per objective (1-3). Such tests are used for general assessment of student learning over broad areas of content. Because of the few number of questions per objective, results cannot be interpreted at the objectives level, but only at the skill cluster level.

At the other end of the content coverage continuum are tests in which a few objectives are tested in greater detail. Usually there are 6-15 questions per objective. Unit tests, mastery tests and **diagnostic tests** are examples of this. Results can be interpreted at the objectives level.

A **norm-referenced test** is one in which scores are available that compare student performance to that of other similar students. Local or "national" norms can be developed for most tests. The term is sometimes used to refer to standardized, published tests, but most tests can be normed. The term is also sometimes used to refer to the method by which the test questions are selected for use on the test -- questions are chosen to "spread" out student scores so that student standing in the group can be determined.

A **criterion-referenced test** is one for which scores are available that compare student performance to a criterion for mastery. Standards for mastery can be developed for any test. The term is sometimes applied to tests which were developed to show instructional progress rather than to compare students to each other. Thus, test questions would be selected for the test to measure important learning outcomes, regardless of whether or not they spread students out in terms of achievement. The term is also sometimes applied to a test development process in which the learning domain to be measured is very carefully defined so that the sample of questions on the test accurately represents the domain.

A **curriculum-referenced test** is one which has been developed specifically to match up with a particular set of curriculum objectives. As such it can be a survey test or a diagnostic test, and it can have norms or have set criteria for mastery. The term mainly describes the degree of match between the test content and a particular curriculum. **We will use the term CRT to refer to curriculum-referenced tests not criterion-referenced tests.**

I. General Information About Multiple-Choice Questions

A. **Definitions**

Stem. The stem poses the problem to the student. It states or implies a specific question.

Correct Response. The correct or best answer.

Distractors. The wrong answers. Distractors should be definitely less correct than the correct response, but plausibly attractive to the uninformed.

Where is the national government of Great Britain located? (Stem)

a. Berlin (Distractor)
b. Birmingham (Distractor)
c. London (Correct response)
d. Paris (Distractor)

Effective Schooling Practices
Related To Assessment

1. Classroom Characteristics and Practices

   1.2 There are high expectations for student learning

   No students are expected to fall below the level of learning needed to be successful at the next level of education.

   1.5 Learning progress is monitored closely

   Teachers frequently monitor student learning, both formally and informally.

   Classroom assessments of student performance match learning objectives. Teachers know and use test development techniques to prepare valid, reliable assessment instruments.

   Students hear results quickly; reports to students are simple and clear to help them understand and correct errors; reports are tied to learning objectives.

   Teachers use assessment results not only to evaluate students but also for instructional diagnosis and to find out if teaching methods are working.

   1.6 When students don't understand, they are retaught

   Teachers reteach priority lesson content until students show they've learned it.

   1.9 Instructional groups formed in the classroom fit instructional needs

   Smaller groups are formed within the classroom as needed to make sure all students learn thoroughly. Students are placed according to individual achievement levels; underplacement is avoided.

   Teachers review and adjust groups often, moving students when achievement levels change.

2. School Characteristics and Practices

   2.2 Strong leadership guides the instructional program

   Instructional leaders check student progress frequently, relying on explicit performance data. Results are made visible; progress standards are set and used as points of comparison; discrepancies are used to stimulate action.

   2.4 Students are grouped to promote effective instruction

   In required subjects and courses, students are placed in heterogeneous groups; tracks are avoided; underplacement is avoided.

2.6 Learning progress is monitored closely

Test results, grade reports, attendance records and methods are used to spot potential problems. Changes are made in instructional programs and school procedures to meet identified needs.

Summaries of student performance are shared with all staff who then assist in developing action alternatives. Periodic reports are also made to the community.

Assessments are coordinated; district , school and classroom efforts work together; duplication of effort is minimal. Assessments match learning objectives.

Staff follow simple routines for collecting, summarizing and reporting student achievement information; results are related to learning objectives. Individual student records are established and updated periodically; group summaries are pulled form individual reports and reviewed over time to check for trends.

2.11 Teachers and administrators continually strive to improve instructional effectiveness

School improvements are directed at clearly-defined student achievement and/or social behavior problems; strong agreement is developed within the school concerning the purpose of improvement efforts.

3. District Characteristics and Practices

3.3 Information about student performance is collected and summarized at the district level. Strengths and weaknesses are identified; reports are prepared and shared throughout the community; special emphasis is placed on progress related to district goals and priorities.

Assessment efforts are coordinated. District-level planning eliminates duplication of effort and ensures quality at all levels; assessments are regular, routine and cause minimum disruption of classroom instruction.

Alignment between tests and the curriculum is checked and improved systematically.

Assessment results are used to evaluate programs and target areas for improvement.

(This information is extracted from *Effective Schooling Practices: A Research Synthesis*, Goal Based Education Program, Northwest Regional Educational Laboratory, 1984; and is based on a presentation by Wayne Neuberger of the Oregon State Department of Education.)

Where Have We Been, Where Are We Going?
Small and Large Group Discussion

Purposes: Participants will know who is working on what, and especially, who is undertaking projects similar to their own.

Activities:

1. Use about 10 minutes to fill out the Testing Information Sheet. Representatives of the same district should fill this out jointly before breaking into small groups. The Testing Information Sheets will be compiled and mailed out to all participants for networking purposes; they will also be used to help tailor the workshop series.

2. Break up into small groups of about 5-6 persons each. Representatives from the same district should split up. There is a colored dot on each name tag. Please join the group with the same color.

3. Use about 30 minutes (5 minutes per person) to discuss local test development efforts in terms of purposes for testing, development procedure, use, development/ongoing cost, benefits/problems, political issues, future plans, etc.

   One person should be designated a spokesperson to report back to the whole group general impressions of the discussion and sources of resources for test development. General impressions could include such things as striking similarities and differences in purposes, what the tests look like, political implications, technical issues, money, etc. The spokesperson does not need to record specific activities from each district because this information will be collected when individuals fill out the Testing Information Sheet.

4. Reconvene in the large group.

Testing Information Sheet

Instructions: Please use this form to describe the nature of your past and future test development activities.

Past Activities:

1. Purpose(s):

2. Subject Area(s):

3. Grade Level(s):

4. Other Features:

Current/Future Activities:

1. Purpose(s):

2. Subject(s):

3. Grade Level(s):

4. Other Features:

Criteria For Analyzing CRT Development Plans

Your plans to develop and use CRTs (or, indeed, plans for any innovation) will succeed based on people. Consider the questions below as criteria to test whether or not your plan will lead to the conditions or changes you want in your school.

1. Does the top management understand and support the improvement effort?

   - o Is there leadership commitment over time?
   - o Are there plans for both directive and emotional support?
   - o Have concerns been identified and plans laid for legitimizing and dealing with them?
   - o Are there few enough innovations in progress that they can all progress successfully?
   - o Will day-to-day incidents support the plan?

2. Is there a felt need by local staff, those who will be involved in and affected by the improvement?

   - o Do the teachers believe the goal is important?
   - o Will some teachers resist the effort?

3. Are the individuals who will implement an improvement involved in planning for initiation and implementation of the effort?

   - o Have steps been taken to inform and involve the whole staff in the planning process?
   - o Does you plan include steps for continuous involvement in the planning processes?

4. Does the improvement fit the local school conditions?

   - o Does you plan consider the effective practices that are currently in place in your school?
   - o Is there room in the plan for individual tailoring? (For example, how results will be discussed by various groups of teachers; how retesting will occur; how standards will be set, etc.)

5. Are there incentives for participation in the improvement effort?

   - o Does your plan contain activities to create and provide incentives for staff members to participate?
   - o Are activities included in your plan to recognize staff members for effective implementation of your prescription?

6. Is there adequate staff development in the plan?

   - o Are specific staff development activities planned?
   - o Will the staff development activities help individuals develop the attitudes, knowledge and skills needed to implement the CRT system?
   - o Are specific activities planned to reinforce and/or reteach important skills over time?

7. Is time allowed for staff to develop the products and practice new techniques?

   - o Is there release time to develop plans and tests?
   - o Is there release time to participate in discussions about results and curriculum/instructional implications?
   - o Do staff members know where to get assistance if they are having difficulty?

8. Will the staff see some results quickly when the plan is implemented?

   - o Will staff be working together differently? More effectively? In more satisfying ways?
   - o Will staff feel that good things are happening?
   - o What else is planned to show some immediate positive results of having CRTs?

9. Are all plan elements ready?

   - o Is the vision of what the testing system will look like and how it will work clear?
   - o Is there an outline of policy that supports this vision?
   - o Is there a plan for implementing the vision? Are the steps clear, well-sequenced and complete? Are responsibilities and timelines clear? Are there adequate resources?
   - o What are the little day-to-day things that will build a perception as well as the reality of success?

(This information is based on a handout prepared by the Science Project, Northwest Evaluation Association.)

The Test Development Process Part 2 -- Test Development Steps

Step 1: Decide on the purposes for testing

Purposes for testing will determine what the tests will look like, who the results will be reported to, how the results will be used, and the resources needed. The handout "Purposes for Testing" illustrates how differences in purposes influence the way the test is constructed. Everyone needs to agree on these purposes ahead of time. This process is heavily "people-oriented." Deciding on the purposes for testing will take more or less time depending on where the district is in the process.

Step 2: Develop Testing Blueprint

The purpose of the blueprint is to specify what the test and test questions will look like. This document guides the development of item pools. The blueprint specifies total test length; the objectives to be covered; the number of questions to be developed on each objective; how objectives relate to each other; and how each objective will be assessed. Ideas for items and specific content is often also included.

Issues are breadth v. depth; specificity of curriculum objectives; how objectives relate to each other; how to decide which objective to assess; how to assess things not easily assessed in objective-format; and reductionism.

The test specifications should be reviewed by all interested parties -- teachers, administration, students? With curriculum objectives in pretty good shape, it should take the committee three to four 2-hour meetings (after training) to development specifications and reviewers about 1/2 day per test to provide input. Add to this the time and cost of word processing, duplication and editing.

Step 3: Develop item pools.

An initial pool of items, prompts, scenarios, etc. is developed for each objective. About 1 1/2 to 2 times the number of items needed is generated. Use preexisting items when possible -- teachers, item banks. Only write when you have to. These items should be reviewed for match to objectives, bias, and technical adequacy.

Item pools are developed by relevant groups of teachers and reviewed by all interested teachers. In the basic skills, for a 200-item pool, it should take two teachers about two days of pulling items and two days of writing items to develop a pool for review. Reviewing a 200-item pool carefully should take a teacher about 1/2 day. Add to this the cost of coding items, entering items, and editing items. These times will increase with other subject areas.

Step 4: Developing pilot test forms

Items marked by reviewers as being of highest priority are pilot-tested. You should pilot test about 125% to 150% the number of items that will be on the final form. The higher number applies to hard to define objectives. At least 50

students per course/grade should take the pilot form. More students are needed depending on other analyses. Pilot testing also requires administration instructions, sample items and graphics.

Teachers should review the pilot forms. Assembly of each pilot form should take 1 day of professional time and 2 to 3 days of support staff time. Administration time varies. Scoring time depends on the type of questions.

Dpending on the proposed use for the results, other special studies will need to be undertaken. For example, HOTS might require having students take the items orally.

Step 5: **Analyzing the results and refining the tests**

There are standard statistics for test items. These, plus teacher comments and observations, can be used to finalize the test items. Basic analysis is very quick with a canned program. Interpretation and deciding which items to keep and which to revise can take 1 to 3 days of professional time per test.

Depending on the special studies and the nature of the use of the test scores, more analysis might need to be done. For example -- test equating, standard setting, bias, relationship to other variables, and factor analysis.

Step 6: **Developing final forms and other documents**

Along with the final forms of the tests, you may want to develop technical reports, user's manuals, a district testing policy, and standard reporting formats. You may also want to start an item bank. Sometimes you will pilot test a second time. Remember that test development is a process not an outcome. Tests are continually refined and revised.

What Do We Need To Get Where We're Going?
Where Do We Need Breakthroughs?

Small and Large Group Discussion

Purposes: District planning time to discuss what is needed to develop CRTs in one's own district; and to discuss the advantages/disadvantages of a cooperative effort.
Provide information to each other on issues and concerns.
Provide information to help tailor subsequent workshops in the series.

Activities:

1. Stay in district groups for this discussion. Each group should have a chartpack and pen.

2. Brainstorm what it would take to get your development effort off the ground; and/or keep it moving along. All responses are acceptable, regardless of level of detail, perceived triviality, perceived silliness, perceived redundancy, or perceived lack of central importance. Needs and concerns could include (but are not necessarily limited to) such things as: resources, information/expertise, equipment, political climate, local commitment, and need for local decisions (e.g., who, what, when, where and why). Concerns could include teacher fears, how information will be used, cost, etc. It is also acceptable to be at the stage where you simply do not know whether this is something you want to do and want more information in order to decide.

3. Prioritize the needs and concerns. List the top 10 concerns and needs. Appoint a spokesperson to report back to the large group.

4. Reconvene back into the large group.

Proposed Homework Assignment From Session 1

Decide if you want to participant in the workshop series.

Participation implies sending a team of 3-5 persons to each session; doing the homework in between sessions; providing the resources needed to complete the project (teacher release time; computer time; etc.) Consider the payoffs as well as amount of time, resources, and commitment needed to develop your own tests. Let _____ know by _____ if this is something you want to do.

If you decide to continue:

- Establish a team to guide the test development process. This team should be able to make decisions about the purpose of the test, subject areas to be assessed, and grades or courses to be included. A good team composition would be 3 to 8 central office personnel, principals and teachers.

- Determine the purpose for your tests. This decision should involve policymakers as well as users of the tests. If different groups have different uses in mind for the tests, potential conflicts will occur down the road.

- Decide on subject matter and grade levels to be included in testing. Consider the condition of the curriculum goals and objectives as well as readiness of the staff in making your selection. The curriculum goals or objectives should be well understood and there should be common agreement as to their importance. In addition, there should be some confidence that the curriculum being taught corresponds relatively well with the written curriculum. If major curriculum work needs to occur, then testing may be premature at this time.

- Establish a teacher review team to help with the test development process. The teachers should be well versed in the curriculum and able to work well with other staff members. The size of the team depends on the scope of the project.

- Let _____ know by _____ purposes, subject area and grades. Also provide the list of names of persons who will attend the workshops. These persons should be a subset of the development team (3-8 persons).

- Be ready to bring to the next session the curriculum goals and objectives in the area to be assessed.

Proposed CRT Test Development Workshop Sequence

Workshop #2: Developing Test Specifications (1 to 2 days)

This session covers purposes for testing, reviewing the curriculum for test development, and developing a blueprint for a test. Participants will begin to develop their testing plans and blueprints. This process will be completed as homework.

Workshop #3: Developing Item Pools (1 to 2 days)

This session assists participants in assembling pools of test questions that match their testing blueprint. NWREL will assist districts in finding item banks to minimize the number of questions that need to be written. Help will be given on writing questions, formatting questions, obtaining input on questions and the logistics of handling large numbers of test questions. Development and review of item pools will be completed as homework.

Workshop #4: Pilot Testing (1 to 2 days)

Participants will develop plans for pilot testing their test questions -- who to test, developing test administration instructions, how to put together the pilot test forms and what other information to collect. Assistance will also be provided on how to score the results and what statistics to generate. Assembling pilot test forms and pilot testing will be completed as homework.

Workshop #5: Finalizing Assessment Materials (1 to 2 days)

Participants will bring their pilot testing results and will learn how to use item statistics to guide item revisions. Assistance will be provided on how to finalize the tests and other testing materials, what to put in a technical report, what to put in a user's manual, what to put in a district testing policy and how to report and use results effectively.

Best Case Scenarios for the Future: Testing that Truly Promotes Curriculum Improvement[1]

Draft Working Paper

Beverly L. Anderson

Education Commission of the States

Denver, Colorado

INTRODUCTION

My main message today is that we need to keep our eyes clearly focused on our testing purposes and their impact on the process of change if we are to achieve the best case scenarios for the future. It sounds so simple. Yet if you go into your typical school, district or state today and ask what their primary testing purposes are, you are likely to leave with a very muddled view of what they are trying to accomplish through their testing activities. If you focused in on how testing impacts the process of change, I would venture to guess that you would find very few people who could give you a satisfactory picture of how various types of tests affect the process most school now are going through—changing to a more effective curriculum.

I would like to first go through two frameworks for describing the purposes for testing—one that is particularly relevant to a local district and its schools and one that is more appropriate for thinking about statewide testing. Then I will look at these testing purposes in relation to the stages of change that seem to describe the process a school goes through as it seeks to make a significant improvement in the way it functions.

I should warn you up front that the conceptualization of these matters that I am about to present derives largely from my experiences from working with many schools and state departments of education on developing or selecting tests and designing testing programs as well as the research I have been involved in for school improvement strategies. I am trying to put these somewhat separate activities together in a way that can help move us ahead in determining what role testing should play in the future. Please do not let my view go unchallenged. The intent here is to force us to think about these matters in new ways and then try them out against the experience of others and the situations in a variety of schools and states. The hope is that this group will generate a significantly deeper and more meaningful way of dealing with the testing issue in the mathematical sciences.

[1] Presentation made at the National Conference on the Influence of Testing on Mathematics Education; June 27, 1986, University of California at Los Angeles, sponsored by the Mathematics Sciences Education Board of the National Research Council and Center for Academic Interinstitutional Programs of the University of California, Los Angeles

A FRAMEWORK FOR TESTING PURPOSES AT THE SCHOOL AND DISTRICT LEVEL[2]

I will just present a framework for testng purposes that I suspect is quite familiar to all of you. I present it now simply to refresh your memory of the multiple uses for testing. I will use it later when we discuss the relationship of testing purposes to the process of school improvement.

Tests are used at the school and district for three major purposes: instructional management, entry-exit decisions and programmatic decisions. Instructional management and entry-exit decisions require test data for each student. Programming decisions can be made based on group data, which allows a sampling of students rather than testing every student.

Instructional Management

Tests play an important role in instructional management decisions. Data from these tests are used for the diagnosis of students' strengths and weaknesses, student placement, and educational-vocational student guidance.

Diagnosis. Perhaps the most frequent use of tests is to diagnose the educational developement of individual students. Here, the teacher is the primary decision maker, although students may also be involved. Teachers often use tests and other performance indicators to assess the student's current development so that the next, most appropriate instructional unit is selected. Tests useful in diagnostic decision making are those that reveal precisely what skills and knowledge the student has or has not mastered.

Placement. If diagnosis determines what instructional units within a course a student needs to master, then placement groups the student according to the next level of instruction best suited to that student's skills. In this case, the decisions are made by administrators, teachers, and guidance counselors who must place each student in the most appropriate course. Math tests, for example, might be used to place students at the appropriate level in high school math course sequence. A test which indicates student ability in math will ensure that students will not be assigned to courses which are too advanced or too elementary for them. Placement tests usually cover a broader range of knowledge and skills than diagnostic tests and are only used once or twice a year. Diagnostic tests may be used on a day-to-day basis. However, completion of grades and courses are also considered in placement decisions.

Guidance. While diagnosis matches the student to an instructional unit, and placement matches a student to a course, guidance can determine an entire program of study. Here, students and their parents assisted by guidance counselors make the decisions. When students decide which educational and vocational program to pursue, they must consider their chances of success and satisfaction. These career planning decisions, typically made in junior and senior high school are assisted by the use of tests that cover broad academic areas and tell the students where they stand in relation to other students. These tests scores can also determine students' strengths and weaknesses which will aid them in making choices. Test scores, of course, should never serve as the sole basis for any guidance decision. The student's academic record, interests and aspirations all merit consideration.

[2] The content of this section was previously presented in Educational Testing Facts & Issues; a layperson's guide to testing in the schools, by Beverly Anderson, Richard Stiggens, and David Gordon, 1980.

Guidance testing, which is generally determined by school or district administrators and guidance counselors, is usually a secondary result of placement or diagnostic testing.

Entry or Exit Decisions

Tests are also used to determine if a student should be placed in an educational program or to determine if a student has completed a program's requirements. For example, tests may be administered in order to select students for programs with limited enrollment (e.g., college entrance or trade school), or to certify minimum competencies (e.g., for high school graduation or occupational licensing).

Selection. The difference between selection and placement is not always clear. Placement, as previously described, groups students in the most appropriate level of instruction. This is an instructional management decision. Selection refers to a process whereby students are screened for admission to an educational program which has a limited number of participants. Admission is based on who is likely to benefit. Here, the key decision makers are teachers and administrators. A test used for the purpose of selection focuses on students' skills and knowledge considered essential for success in the program, and compares students' relevant skills and knowledge so that those most likely to succeed are identified. Admission to college or into a particular course (for example, airline pilot training) are prime examples of selection. However, test scores are not the sole basis for selection decisions. Previous academic record and other performance criteria may also be considered.

Perhaps, the most common use of selection testing is the college entrance examination. Colleges require a specific entrance examination and interested students register with test publishers who carefully control the administration of the tests at various locations across the country.

Certification. Tests often play an important role in certifying acceptable minimum levels of educational development in students. For example, a teacher might use a test to certify mastery of beginning verbal skills required for completion of a certain course. Or, a district administrator applying Board of Education graduation standards might use an examination in order to test a student's mastery of minimally acceptable skills. Or, members of a certain technical profession might use a test to certify competence in that profession. Since, in each case, those taking the exam must pass the test to be certified, the test must focus specifically on clearly stated minimal competencies.

Programmatic Decisions

A third of use of tests is to assist in program planning. In this instance, test data may be helpful in providing the basis for developing a new program, allocating funds or evaluating existing programs. Such testing falls into three categories: survey assessment, formative program evaluation and summative program evaluation.

Survey Assesment. A common use of testing in education is to survey student achievement and analyze trends over time in order to assist in program planning. This kind of testing is usually designed to raise issues for further investigation. For example, the test results might prompt such questions as, why are math scores gradually declining in the district (or state or nation)? Or, why are reading scores of fourth graders consistently below national averages while those in other grades are above average? The test data are used to identify which aspects of the educational system need to be more thoroughly investigated as well as possible reasons for unsatisfactory

performance. For this purpose, achievement test scores— sometimes from random samples of students— are gathered annually, then averaged across the entire school, district or state, and used to indicate the level of student development. In order to show trends, test scores are frequently compared from year to year. This information then becomes a basis for setting educational policy and allocating funds. Typically, educational administrators are the primary decisions makers, but they must justify these decisions to the ultimate decision maker, the taxpayer. Tests used to assess an educational program must cover broad content and skill areas in order to provide valid information for program changes.

Formative Evaluation. In formative evaluation, the goal is to determine which instructional units or features of a specific educational program (e.g., remedial reading), are effective and which need revision. In this instance, tests are used to measure what the students learn in a specific program and the results are used to help shape or revise the program during its formative stages.

Summative Evaluation. Summative evaluation reveals a program's overall merit, and suggests whether or not a program should be continued, terminated, or expanded. Tests designed to assess knowledge gained from a program are an important part of such an evaluation. Teachers, program, building or district administrators, and the public, represented by the board of education, may be involved in summative evaluation decisions. Tests may be given both before and after instruction, with retesting after an interval to determine the student's retention of knowledge.

A FRAMEWORK FOR TESTING PURPOSES AT THE STATE LEVEL[3]

Consider three purposes that seem entangled in state student testing activities- monitoring performance trends, ensuring accountability, and advancing the curriculum. Although these purposes are not totally distinct, I would like to consider them separately and explore the differences in test design, administration, and reporting that flow from each purpose. Table 1 provides an overview of these relationships.

Monitoring Performance Trends. The initial purpose of most statewide testing was simply to observe learning trends. The emphasis was on "Where are we?" not on "Whose fault is it that we are where we are?" Unfortunately, the assignment of responsibility came so quickly that some people forgot that the objective observation of trends needs to continue and that it implies different test characteristics that an accountability test.

Monitoring tests need to be broad based. That is tests need to cover very basic skills required of all or nearly all students, higher order skills, and skills in a broad range of subject areas. The state (or nation) as a whole needs information on these important skills.

Accountability. Tests used for accountability may well need to be structured differently than tests used for monitoring. Tests used for accountability need to be based on widely agreed on learning objectives. Who is accountable for what needs to be clear before testing occurs. Tests need to be administered with a high regard for security and fairness; analyses need to show what specific improvements in learning are

[3] The content of this section was previously presented in my article "State Testing and the Educational Measurement Community: Friends or Foes" in Educational Measurement Issues and Practice, Summer, 1985.

needed. If students are held accountable, for example, then scores need to be reported to individual students in ways that make it clear to them what improvements are necessary. If district scores need to be reported both to districts and to those who hold districts accountable (e.g. the public, the state education agency). Test results need to be analyzed and reported in ways that show which deficiencies need correction. Of course, seldom is only one party accountable. If a student is accountable for learning, the teacher is almost always responsible for adequate instruction and the principal and district are responsible for supervising and guiding instruction.

The relationship of testing to curriculum requirements is an interesting one. If states test solely to monitor student trends, one could argue that whether states or districts are responsible for defining the curriculum does not particularly matter. Nor does it matter if all skills monitored are widely taught. The very tricky task in this situation is, of course, for a state to report trends in ways that do not put inappropriate pressure on districts to change what they are teaching (e.g., do not compare districts or classrooms). If a state tests for accountability, however, the relationship between curriculum and test becomes very important. Take the Florida situation. The state establishes specific skills that districts are to teach, tests measure those skills, and results are used to hold the districts accountable. Reporting district comparisons and using public pressure to focus schools on required skills is consistent with designated responsibility. But if a state- either by constitution or tradition- leaves curricular decisions to districts, there is an inconsistency between state testing for accountability and district curricular responsibility. Such situations point to the importance of state leaders, as well as the measurement specialists involved, being clear on whether the tests are for monitoring or accountability and, once the purpose is established, being sure test design, administration, and reporting are consistent with that purpose.

Curriculum Advancement. In one sense, testing for accountability is actually the opposite of reform, in that it focuses on aspects of the curriculum regarded as unchanging essentials. Testing for curriculum advancement, on the other hand, focuses on change, not constants. Testing has played a major role in two recent instances of curriculum advancement. In the early 1970's, NAEP attempted to measure writing skills, knowing that this would be extremely expensive and that scoring techniques were still inadequate. The result has been tremendous advances in the teaching and testing of writing. About 3 years ago, NAEP reported that students were gaining in certain basic skills but losing in higher order thinking skills. Several states are now attempting to measure higher-order skills, using testing as a means to refine the curriculum.

Tests used primarily for curriculum advancement will look very different from those used for accountability. They may not necessarily be based on widely agreed on learning objective. They may be designed to measure skills that have not yet been clearly defined. Security is a less critical issue, because no one person or group will suffer consequences based on test results.

By the way, you may be interested in a little historical information. Before 1970 there were only 6 states in this country that had state assessment and none had what we now know as minimum competency testing. During the '70s, state assessments grew phenomenally, and then minimum competency testing emerged in full force in the mid to late '70s. We are now at the point, that, according to a recent survey we did, there are only two states in the country that do not have one or both of these types of testing. So within a 15-year period, state testing has changed from an infrequent characteristic of state departments to a near universal one.

THE CONTEXT FOR IMPROVEMENT

Before I talk specifically about the stages of change in the school improvement process and its relationship to testing purposes, I would like to take a few minutes to focus our attention on future conditions that are likely to exist in this country that will dramatically impact on the education system. I will comment on changes in the economic conditions, technological changes, demographic changes and some changes in our students and teachers.

First of all, in regard to the economic conditions, one of the things that you well know is happening in this country is that we are losing many of our lower level jobs. Industries are capitalizing on the fact that they can, for example, do assembly work more cheaply in other countries or through automation. We are also faced with the dilemma that more and more people with fairly high levels of education — are not being able to find positions that fully utilize their skills and knowledge.

These economic changes are putting at least two kinds of pressures on the education system. First, there is the pressure to better match education and training to the types of positions that will be available in the future. But there is a second pressure. People are increasingly realizing that education must prepare students not only to be productive workers but also to be productive citizens and people who can enjoy life and obtain value from things other than just economic gain.

A second characteristic of the future will be increasing technological changes. I'm not going to dwell on this one because I am quite sure you are all well aware of the tremendous impacts of these changes.

Changing demographics are a third category of change that I do want to emphasize for a moment. Slowly but surely deeply significant changes are occuring in the demographic make-up of the country and in the cultures that are being brought into the schools. Let me read a couple paragraphs out of the May 14, 1986 issue of Education Week.

> Today we are a nation of 240 million people, about 50 million (21%) of whome are Black, Hispanic, and Asian. Although federal and private projections vary, they all point in the same direction. Soon after the turn of the century, one out of every three Americans will be non-White. Immigration patterns and differential fertility rates among various groups are significantly changing the nation's racial composition.

I could go on to read you many more statistics about these changes, but the general jist is that we need to be understanding different cultures, recognizing that schools will increasingly have a different mix of backgrounds and that these changes have significant implications for how we go about testing and teaching in the schools.

Another major social change occuring in our nations relates to the changes in family units and the sense of community that exists within various groups. Another paragraph from the same article reads:

> Next September, more that 3.6 million children will begin their formal schooling in the United States. One out of four of them will be from families who live in poverty. Fourteen percent will be the children of teenage mothers. Fifteen percent will be physically or emotionally handicapped. As many as fifteen percent will be immigrants who speak a language other than English. Fourteen percent will be children of unmarried parents. Forty percent will live in a broken home before they reach 18. Ten percent will have poorly

educated, even illiterate parents. Between one-quarter and one-third will be latch key children, with no one to greet them when they come home from school, and a quarter or more of them will not finish school.

What a fertile ground for producing "disconnected" or "at-risk" youth. And what strong implications for the environment of the school.

A final set of changes—those in the teaching force. ECS, along with many other groups this year, has been particularly involved in looking at how to raise the sense of professionalism among teachers, how to make teaching a more attractive field and how to improve the working conditions in the schools so that highly qualified people will once again be drawn to teaching.

When I put all of these changes together in my mind, I see tremendous implications for what we do in schools. We need a more integrated curriculum; we need more attention to problem solving and critical thinking skills; we need more personal and wholistic attention to the needs of children; and we need to understand what will attract and return good teachers. All of this also says to me that we can not go about piecemeal change any longer. These matters are much too interconnected. We need to become far more sophisticated in understanding the process of change and how levers such as testing play very significant roles in facilitating or inhibiting effective change.

LINKING TESTING TO THE STAGES OF THE SCHOOL IMPROVEMENT PROCESS

What I'd like to do for the next few minutes is to take some information from a study we did recently at ECS on how school improvement operates in exemplary schools. We went into 10 states, 40 districts and about 150 schools that were engaged in improvement efforts that seemed to be working reasonably well. What we tried to do was to identify the stages in the change process and the characteristics of each stage. If you looked at a list of characteristics you would see many familiar items. That's one of the deceptive things about change. We all know how the process works in a broad sense. What we need to understand is more clearly the sequence of events and which are the most powerful characteristics that inhibit or facilitate change. With this awareness we can then reflect on the impact of different types of testing arrangements at each stage.

The first stage of change—one that's often overlooked as a definite stage—is the initiation phase. We are currently initiating change in nearly every aspect of the education system. All of this could come to naught if we don't look at what we're really trying to do during that initiation phase. During that time, the critical thing is that we're making people aware of the discrepancy between what is and what should be (in the sense of what's going to work in the future). What we're constantly trying to do at this stage is raise people's consciousness of that gap and get people to buy in and say, "That's serious. We've got to do something about that." Particularly, it needs to happen among leaders so that they can be in a position to move things along.

Some analyses done of the NAEP data several years ago illustrate the use of testing at this stage. What we did was to take the data in reading, math and science, and look at the different trend lines over several years for low performing students and high performing students. Basically what we found in several of the subject areas was that the lower performing students were increasing their skills, whereas the higher performing students were either holding their own, or actually going down hill. Then we looked at it another way. We found that most of the increases of the lower students were in low level skills, not with the higher order thinking skills. The high performing students, likewise, were not improving in higher order thinking skills. These analyses conveyed the

message: "Look, there are discrepancies here that we just can't live with. We may be doing OK in lower level skills, but we need to do something about the higher order skills." Thus this testing information promoted and helped focus the type of change needed in the initiation stage.

Once there is agreement on the problem among a decent number of people, particularly the leadership, the change process moves into the initial implementation stage--a critical time. As we looked at improvement programs around the country, we found several important characteristics at this stage. First, during that time you need to create an environment of collaboration and shared responsibility. You also need to have an environment that creates a common understanding and a shared vision of where we are going. There needs to be time for people to try new things, to practice, to work things out, to go in this direction for a while, then regroup and move in a somewhat different direction. People are exploring options in a collaborative team sense.

One point that should be evident from the comments this morning is that right at this crucial time of change, when we need to be opening up the system, tests are often closing it down. If we're going to promote this initial implementation stage, we need to be clear on what kinds of tests will do that. A good example of the type of test that promotes this open exploring environment is the approach used in directing writing assessment. Let me just make a few points about that for those of you who are not familiar with it.

The whole notion behind the direct measures of writing assessment is that you ask a student to actually write, and then you score that paragraph in some fashion. There are two critical features to this approach that I think are important. First of all, the students are actually demonstrating a skill that you want, and secondly, you're using the scoring methods to promote learning on the part of teachers. I'll just use an example from a writing assessment project we were involved in when I was back at the Northwest Lab in Portland. 9th graders were given this challenge:

> The school board is about to pass a policy that we have year-round schools. Your task is to write to the school board and give them your opinion. Provide support for your point of view.

The 9th graders wrote letters, then groups of teachers got together to score them using one of several methods, all of which require debate and discussion among the group. In these scoring sessions, teachers are hearing from one another what each one considers to be the important elements in writing. There is a lot of learning going on about how to teach writing.

Basically, there are 3 different ways to score these direct measures of writing. One is the "wholistic" approach. Here the scorers read the paper and give an overall impression of the paper. They usually give papers a score from 1 to 4. A second way is an "analytic" scoring approach where you actually break out the components of writing, like the organization of the paper, the wording, the ideas. You score for each of these components using set criteria. Again, these are debated and discussed. The third scoring approach, "primary trait" requires that you score the paper based on the purpose for writing.

My suspicion, (and I think some of you are already doing this in this room), is that this technique or style of testing is one that can be easily transferred and used in problem solving and critical thinking situations. You could just as well have given a student a situation where a landscaper had to calculate the amount of sod he needed for a rather unusually-shaped lawn. Then you ask them not only provide the answer but also to

explain their process of arriving at the answer. You could use a scoring approach parallel to that used in writing and end up training teachers at the same time.

I use this simply as an illustration of what testing should look like it if is to promote the initial implementation phase. You need to give teachers room to learn while you are using testing to zero in on what the skills are that should be taught and measured.
Now let's look at the third stage of change, which we call complete implementation. You now have a fairly clear focus on the skills to be taught, have materials and techniques refined. You are now ready to have all the appropriate teachers involved and are expecting them to master the teaching process. The type of test you need at this stage is one that has a sharp focus on the essential skills. Whether it is open ended or multiple choice isn't the issue. Rather the concern is whether the test clearly measures the essential skills. Multiple choice tests are likely to be very appropriately used here to measure many skills.

The fourth stage of the change process is referred to as institutionalization. Here's where you make structural changes or role changes that stabilize and support the new curriculum (or other change). Given our changing environment, you don't want to stabilize everything, but you want to find a few key elements you want to keep in place and then have a system that allows other things to move. At the institutionalization stage, one of the important testing issues is that local and state testing need to be compatable. One of the things we're learning about school improvement efforts around the country is that many schools are able to get improvement going, often because of a fairly charismatic principal. Once that person leaves, or other people just get tired, things start to die unless everything up the line, in terms of the regulations coming from the state, or the guidelines from the superintendent, are in support of that new situation. Thus it's important that we get our testing systems working together across lines, and working with other structural and regulatory features.

I want to draw something on the board for you (Figure 1). Take these four stages and go back to the three purposes for state testing: monitoring, accountability and curriculum advancement. You'll see that typically what happened across the country is that we went from monitoring to accountability and then to curriculum advancement, (if at all) which is in the wrong order when you consider the stages of change. We started out OK. We needed the monitoring to get people to see discrepancies (at the initiation stage). What we needed next were curriculum advancement type tests to nurture that developmental and exploratory time of initial implementation of change. Way at the end is where we needed accountability testing.

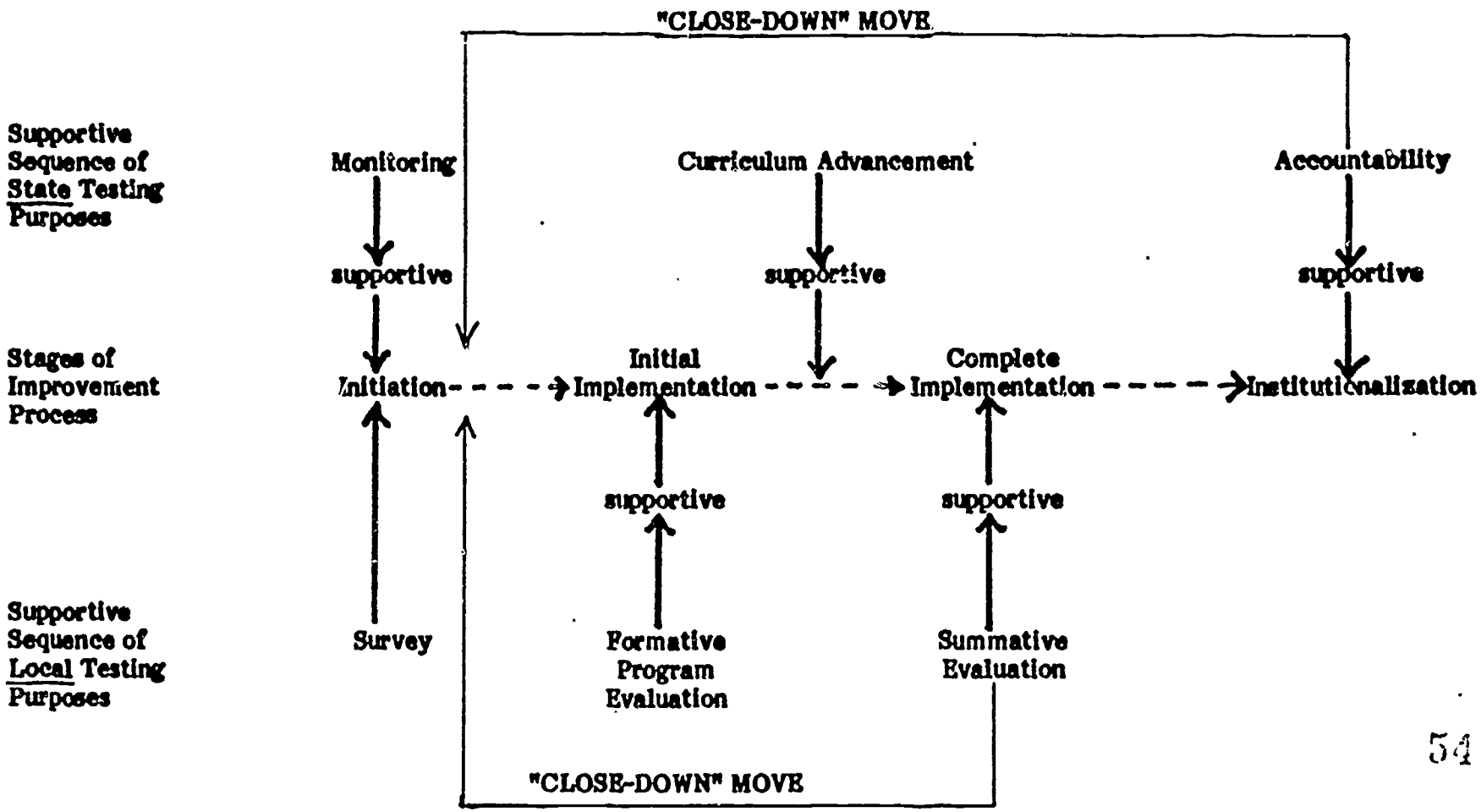
Likewise look at the local testing purposes in the programmatic decision making category of our first framework. Survey testing is first. Then testing needs to be joined with broader programmatic evaluation techniques and used in a formative way to shape the new program. Finally summative evaluation is needed.

SUMMARY

Testing, is a tremendously powerful tool. As we seek to make major changes in our educational system, we need to be sure that we're using tests in a way that doesn't close down change, but rather open the system and allows the exploration and creative development of new approaches before we hold people accountable.

FIGURE 1

Aligning Testing with the Stage of the Educational Improvement Process



= testing purpose used as a support to effective change

= testing purpose at a stage when it closes down the system

= desired progression of change

HARD COPIES OF TRANSPARENCIES

CURRICULUM REFERENCED TEST DEVELOPMENT SERIES

INTRODUCTORY WORKSHOP

DEFINITIONS

Standardized

Survey Test

Diagnostic Test

Norm–Referenced Test

Criterion–Referenced Test

Curriculum–Referenced Test

IMPLICATIONS

1. Different test types serve different purposes
2. Therefore, users need to have purposes clearly in mind before beginning test development
3. No single test is likely to serve all purposes
4. A good testing system is a combination of NRT and CRT

WHY HAVE CRTs?

1. The place of assessment in effective schools

2. Diagnosis and instructional planning

   The commitment that all students can learn

3. Objective information

4. Motivation

USE OF ASSESSMENT IN EFFECTIVE SCHOOLS

- Individual student progress monitoring
- Grouping
- Monitoring status of educational program
  - Accomplishing goals
  - Spotting potential problems
  - Adjusting curriculum
  - Targeting program improvements

PURPOSES FOR TESTING

Instructional Management

- o Diagnosis
- o Placement
- o Guidance

Entry and Exit Decisions

- o Selection
- o Certification

Programmatic Decisions

- o Survey assessment
- o Formative evaluation
- o Summative evaluation
- o Performance trends

Accountability

WHERE HAVE WE BEEN? WHERE ARE WE GOING?

1. District groups complete Testing Information Sheet.

2. Move to discussion groups according to the colored dot on name tags.

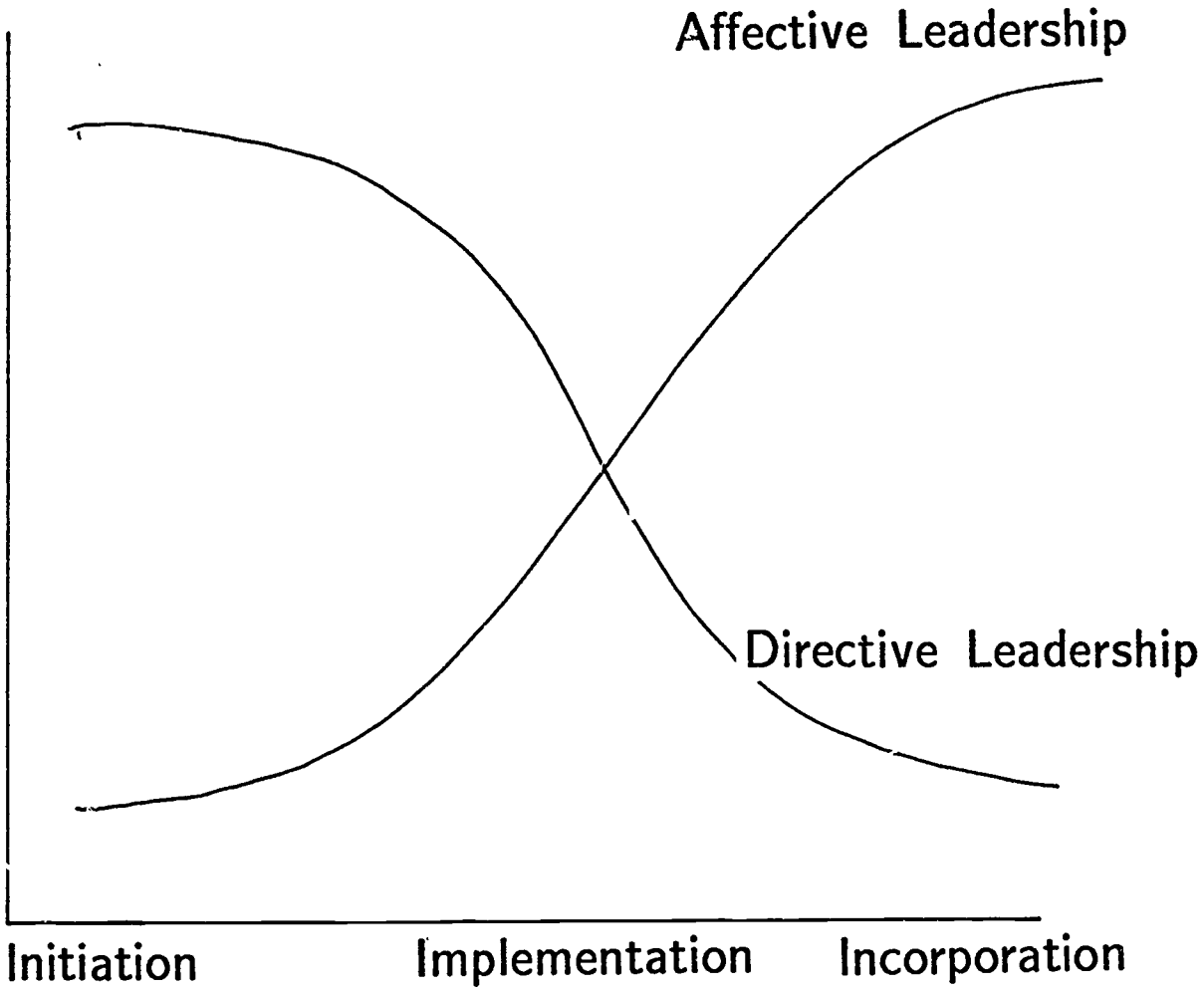
3. Discuss local test developent efforts -- purposes for testing, development, use, costs, benefits, problems, political and social issues, future plans, etc.

4. Appoint spokesperson to report back to entire group.

5. Reconvene large group.

WHAT WILL CAUSE THE PLAN TO SUCCEED?

1. Administrative support
2. Felt need by staff
3. Implementers involved in planning
4. Plans fit local conditions
5. Incentives to participate
6. Adequate staff development
7. Adequate staff practice time
8. Quick results
9. Plan is ready


Affective Leadership
Directive Leadership
Initiation
Implementation
Incorporation

IMPLICATIONS OF ADEQUATE ADMINISTRATIVE SUPPORT

- Commitment to long–term effort
- People need to know this is where attention will be focused
- Do not do too many things at once
- Lack of administrative support reinforces negative perceptions
- Time and resources
- Incentives
- Pay attention to incidents

WAYS TO BUILD FELT NEED

- Formal needs assessment
- Informal discussions
- Look at data
- Quickly find a way that data saves time
- Involve users in planning

TEACHER CONCERNS ABOUT TESTING

1. How will the results be used?
2. Will CRTs restrict the curriculum?
3. Will the curriculum be trivialized?
4. Will there be less freedom for teachers?
5. Why should I spend time on this?
6. Is there release time for doing this?
   Resources?
7. There is already too much testing.

TEST DEVELOPMENT STEPS

Step 1: Decide on testing purpose(s)

Step 2: Develop testing blueprint

Step 3: Develop item pools

Step 4: Develop pilot test forms and pilot test

Step 5: Analyze pilot test results and refine tests

Step 6: Develop supporting documents and systems

WHAT DO WE NEED TO GET WHERE WE'RE GOING?

1. District group brainstorming —— resources, information, expertise, equipment, climate, political concerns, local commitment, need for local decisions, misuse of results, cost, etc.

   All responses are acceptable.

2. Prioritize needs and concerns. List the top 10.

3. Appoint a spokesperson to report to the entire group.

4. Reconvene the entire group.

PROPOSED HOMEWORK FOR SESSION 1

1. Decide if you want to participate.
2. If so ——
   - o Establish a developent team
   - o Determine purpose(s) for testing
   - o Decide on subject and grade levels
   - o Set up a teacher review team
   - o Bring curriculum goals and objectives to session #2
   - o Communicate with Ray or Judy

PROPOSED WORKSHOP SERIES

Workshop #2: Planning for test development

Workshop #3: Developing test questions

Workshop #4: Pilot testing

Workshop #5: Revising questions, using results, supplementary products

CURRICULUM–REFERENCED TEST DEVELOPMENT WORKSHOP SERIES

#2: Developing Test Specifications

PRESENTOR'S OUTLINES

73

Introduction To Workshop #2
Presenter's Outline

Purpose: To describe the purpose of the workshop and go over the agenda.
To introduce participants to the content of a prototype "test plan" which includes all the decisions made so far about the test(s) and includes the test specifications. This document will be reviewed by interested parties prior to the developed of the test(s) themselves.

Materials: Agenda
Test Plan Report Handout and Transparency

Format: Lecture/discussion

Time: 15 minutes

Points To Make:

The most important part of putting tests together is planning what they will look like, why the tests are being given, logistics of testing, etc. Careful planning simplifies the task of assembling the tests and helps to avoid problems later.

This workshop and the associated homework is intended to result in a testing plan that will guide the test development, document the decisions made and allow interested parties to provide feedback before the actual tests are assembled.

Go over the agenda.

The prototype Test Plan has three parts -- background information, clarification of curriculum objectives and how the time available for the test will be distributed among content and item types. The main focus of the workshop is the latter two items. The first part provides an opportunity to summarize decisions made so far about the test.

Go over the sections of the report entitled "Purposes for testing," and "Other considerations..." Point "d" under "Other considerations" can include such things as who will be tested, where students will be tested, language of testing, overall testing philosophy, time or money constraints, etc.

Participants can make notes on what they will include in these sections of the Test Plan as appropriate.

Test Specifications Part I: Clarity
Presenter's Outline

Purpose: To provide an overview of what test specifications are and why they are important.
To give participants practice in deciding the level of detail they need in their test specifications.

Materials: Test Specifications Purpose and Description handout and transparency
Clarity Considerations handout and transparency
Examples of need for clarity in objectives, transparency
Samples of Test Specifications handouts (Glendale; Grade 6 CBM; Utah Science; Lewiston)
Bloom's Taxonomy
Overhead projector & screen

Format: Lecture/discussion/large and small group activities

Time: One to two hours

Content of Presentation:

Define Test Specifications (Test Blueprint)

The two sections of the Test Plan that we will emphasize today are your specifications (blueprint) for the test. These outline exactly what the test will look like in terms of content and length. Content covers such things as how each objective will be measured, clarity on the exact content covered by each objective, and various approaches to assessing the content. Just exactly what this entails and the need to specify will become clear as we go though examples. Length includes total test length and the relative numbers of questions that cover each objective and skill cluster -- in other words, how the total time for the test will be divided up among objectives and tasks.

We will discuss clarity issues first, and then talk about length.

Why There May Need To Be Additional Clarity in Objectives Before Tests Can Be Developed

Many times, curriculum objectives are not written with the clarity needed in order that everyone has a common understanding of what is to be taught and assessed. For example, consider the following objectives:

a. "Know that the sum, difference, and product of integers are always integers." What does "know" mean? How will students act when they "know?" Restate the definition? Pick out examples of the principle? State the principle in symbols? Cite this principle as an axiom when doing proofs?

b. "Explains means of preventing, curing or controlling diseases and conditions." What is appropriate for grade 2? Grade 4? Grade 12? Explain in writing or orally? Recall of information only? What would a teacher teach? What would one assess?

75

c. "Students will differentiate among forms of propaganda in written selections." What forms of propaganda for grade 5? Will students be able to apply titles to types? Is this all they should be able to do?

We are attempting to build curriculum-referenced tests. The purpose of such tests is usually to be able to infer the degree to which students know information, or are able to do various things. From the sample of questions on the test, we want to infer the degree of mastery of a domain of skills. Therefore, in order to be interpretable, the domain from which we will select representative items needs to be clear.

The point here is that if we are to know what to teach and what to assess we need to have a sufficient degree of clarity. More clarity than is required takes up unnecessary development time. Less clarity leads to confusion and inability to interpret the results. We will spend some time trying to build experience with knowing what is the necessary degree of clarity.

Things To Look For When Deciding If Objectives Are Clear Enough

Your objectives need to be clear enough so that the persons who find or write test questions have the same interpretation of what should be tested as the persons who wrote the curriculum. Also will future teachers interpret them the same? Are the interpretations clear enough so that when the results are used to profile student strengths and weaknesses you will know what to do about it? Some questions to ask follow. Have participants look at their own objectives as each point is raised to see whether any of their objectives might need to be clarified.

1. Is it clear what cognitive level is being assessed? Is it recall of facts? Higher-order thinking skills? For example, the "know" objective cited above. This might be assessed, and taught at various cognitive levels. Bloom's Taxonomy can be used to specify levels. There are other classification schemes they can use.

2. Is the objective clear enough so that if a group of appropriate teachers look at it they would agree on what to teach and what test questions would align? For example, "The student will read a word problem and recognize from a list of equations the correct equation used to arrive at the solution." How hard will the equations be? Symbols or numbers? Equations they have encountered before or new ones? Should they also be able to write equations as well as recognize them? Find instances in daily life where they need to generate an equation to solve a problem?

   Is it clear all the things that could be included in instruction and assessment that would be included under this objective? For example "Measure length, volume, mass, temperature, and time using the metric system." Does this also cover reading values, comparing values, converting units and interpreting results? Should they also be able to choose the appropriate measuring device and appropriate units of measurement? Estimate?

   One problem with very specific specifications is that sometimes they become overly restrictive. For example, Glendale, revised CRT Model p. 3 (also listed under b above). You need to be careful that you represent everything that is appropriately covered by teachers for each objective. Elaborations of skills such as HI CBM 6 and Glendale, Old Format may be useful.

Overly restrictive interpretations can lead to teacher conplaints of "doesn't measure what we teach" and to reductionism. Need to sample from the entire domain of interest.

3. Are there any aspects of the stimulus or response that need to be clarified. For example, "Identify the place value of a digit in a given whole number." How many digits is the maximum length? Do you want to specify the most common types of errors as distractors?

4. How will objectives fit together across grades or across subjects in the same grade? Many times scope and sequences repeat objectives across grades. For example "Students will recognize the long vowel sounds" may be introduced in K, elaborated in grade 1 and mastered by grade 2. Will this be assessed the same way in all grade level? Are there some vowel sounds to be covered in K and others in later grades? Will questions increase in difficulty as grade level increases? If so, how? Will students be tested individually in K and in groups in grades 1 and 2?

5. Are objectives repeated across subject areas. Some examples are graph reading in math, social studies and study skills; letter sounds in both reading and language arts; and writing in language arts and subject areas. Are there clear reasons for these repeats? How is instruction different? How will the assessment be different? How will instruction and assessment complement each other to give a full picture of all aspects of repeated skills? Should repeats be moved to just one area?

6. Are there any holes in the curriculum? Are there important skills that aren't covered anywhere? Do the objectives reflect current thinking about what expertise in a content area is? A curriculum-referenced test can only be as good as the curriculum.

Review of Sample Test Specifications

Go over the various samples and review them for:

- o Clarity
- o Overrestrictiveness
- o What amount of detail is enough? Which have more detail than is needed (and are therefore a waste of time)? How much detail is just enough?

The samples are in order of formality. You don't necessarily need to be formal to get the level of detail you need. Use the rule of thumb: "Will teachers agree on the range of knowledge, skills and performance that is implied by the objective?"

Emphasize that the test specifications do not need to be perfect the first time. The process of development, review and using the test results will assist in refining what you want.

Review Of Their Objectives

Use participant objectives as examples. Remind participants that observations on their objectives are not meant to imply that they are not good. They are probably fine for the original purpose. But test development requires looking at them in a new light.

Do some of their objectives as a group and then have individual work. Keep going until they have the idea of what amount of detail is the right amount -- not overboard and not too little; just until they are reasonably assured that everyone will have the same interpretation.

Objective Specifications Exercise (Optional)
Presenter's Outline

Purpose: To provide an experience in developing very formal test specifications for one of their own objectives.
To help participants clarify what level of detail is just enough.

Materials: Specifications Form handout and transparency
Exercise Instructions handout and transparency
Blank transparencies
Overhead projector and screen

Format: Small and large group discussion

Time: One hour

Content:

We want them to do a very formal test specification so they know what's involved and so they get a feeling for what is important to specify. We do not recommend that they do these for all their objectives because (a) they probably don't need to, and (b) they would never get to anything else. But, they do need to know when they can leave objectives as they are and when they need to expand them.

The exercise has the following steps:

1. (10 min) We will do one as a group using the UT science approach to clarification.

2. (20 min) Then they will work in district groups. Each group will pick one to three objectives to use to write a formal test specification. They should pick objectives that can be assessed in multiple-choice format. Two to three persons will work on each objective. Use the attached objectives specifications form. They can use any of the more structured test specification samples as examples of the types of things to include.

3. (20 min) Review and critique each other's efforts.

4. (10 min) Report back and discuss. Especially discuss what they discovered about the level of detail required and other issues/concerns. Be sure to emphasize that we do not necessarily recommend this level of detail on all objectives -- only when it is needed to clarify what items will look like. Also emphasize that they not need to have them perfect the first time. Refinements will occur as they proceed.

Test Length and Distribution Among Objectives
Presenter's Outline

Purpose: Provide participants with rules of thumb for estimating test length, identifying how objectives should be measured, relating test length to uses, and deciding how to distribute items across objective.

Materials: Test Length and Distribution Among Objectives handout and transparency
Sample objectives from Grade 6 CBM, and Lewiston.
Overhead projector and screen

Time: One hour

Content of Presentation:

Go through each of the steps in the handout. At each step outline the issues and rules of thumb and then have the group apply the concepts to examples in Lewiston, Hawaii or Utah. Lewiston is good for all steps and the case where the same number of items will be written for each objective. Utah is good for the example of different numbers of items for each objective because of the uneven levels of detail at which the objectives are written.

Points to cover are:

1a. How to determine which format to use for various objectives. (Go through examples in handout and then use Lewiston for deciding on multiple-choice v. performance. Use Grade 6 CBM to look at teacher and student rates.)

Some objectives could be slightly rewritten so that they can be measured in objective format (see Lewiston). They will have to use their own judgment as to whether such rewrites maintain the spirit of the objective.

1b. Relating number of items to uses. For example, need at least 5-6 items to report separately. This can be at the individual objective or cluster levels. (Go over examples in handout and then use Lewiston as an example.)

It is OK not to have the same number of items per objective. Some are written to cover more ground. E.g. "provide definition" v. "add, subtract, multiply and divide fractions." (Go over example in handout and then look at Lewiston.)

2. How to estimate amount of testing time. (Go over rules of thumb in the handout and apply this to Lewiston example we have been developing.)

3. Modify plans as needed in order to balance testing time against information obtained. Modifications can include: combining objectives/clusters; reporting at the cluster level rather than the objectives level; only assessing some of the objectives; testing more objectives in structured format; etc. (Go over examples in the handout then get their ideas on how/whether our running Lewiston example needs to be modified.)

It is a good idea to summarize the numerical coverage. See HI as example.

They will produce about 1 1/4 to 2 times as many questions as they will need on the final forms.

Remember that this does not need to be perfect the first time. Give it your best shot. Teacher reviewers of the test specifications will point out additional things. Giving the tests will indicate other things. You will refine what you want as you go along.

Reviewing Objectives
Presenter's Outline

Purpose: T ...icipants go through the objectives for one of their courses and indica... each: the testing format, whether clarity needs to be added; the number of items on the final test, and the number of items to be included in the item pools.

Materials: Review Task handout and transparency
Hawaii HSTEC handout
Test Plan Report
Overhead projector and screen

Time: One to two hours

Exercise Tasks:

1. Look at each objective and indicate:

   a. The required testing format -- multiple choice, matching, T/F, essay, performance, teacher rate, self rate. Some objectives can be altered slightly to make them suitable for testing in multiple-choice format. (Do this when it is judged that the spirit of the objective is not changed too drastically.)

   b. Whether the objective needs more clarity. Some of these may be easy enough to clarify now. Others may need more work at home.

   c. The ideal number of items or performance situations that would be required to get a good feel for how students do.

2. Estimate the time required to give the ideal test.

3. Modify your plans to accommodate testing time. Modifications may include combining objectives, only generating scores at the cluster level not the objective level, only testing certain objectives and/or modifying more objectives so that they are testable in structured format.

4. List the number of items to generate for the item pools. This may be overall for the entire objective or, if specifications are listed under the objective, for various aspects of the specifications (see Hawaii HSTEC).

5. Develop summary table of item numbers. (See the Test Plan Report and Grade 6 CBM for samples.)

If participants have to choose between getting greater clarity on some objectives or making it through all the steps above, have them get through all the steps. It may be hard to estimate the numbers of items needed if an objective is very unclear, but this probably will happen only rarely. Just mark the objectives needing further clarity and proceed with assigning numbers.

Test Plan Report Outline
Presenter's Outline

Purpose: To provide participants with an outline of what should be provided to teachers who will review the test specifications.

Materials: Test Plan Report Outline handout and transparency
Letter To Teachers handout
Grade 6 CBM example
Overhead projector and screen

Format: Lecture/discussion

Time: 15 minutes

Content of Presentation:

Emphasize the need to have teacher teams review the blueprint. This builds ownership as well as helps to ensure that the tests will cover what you want them to. A sample cover letter to reviewers is in the handouts.

Go over and discuss each part of the Test Plan Report as needed. Parts have been discussed before. They can feel free to modify any part they feel is necessary.

An example of how this played out for Grade 6 CBM is attached. This points up the need to have clearly in mind what you want to do so that you can gather the relevant information from users.

Wrap-Up For Workshop #2
Presenter's Outline

Purpose: Plan Workshop #3

Materials: Homework handout and transparency
Proposed Session #3 handout and transparency
Overhead projector and screen

Format: Discussion

Time: 15 minutes

Topics to be Covered:

1. Assessment of workshop #2. Was there time enough to cover all the topics? Should it be two days? Did they get enough information to complete work at home?

2. When should workshop #3 be? They probably need at least two months to finish test specifications and have them reviewed.

3. Review the homework. Adjust it as needed.

4. Review the content for workshop #3. Adjust it as needed.

PARTICIPANT HANDOUTS

DEVELOPING TEST SPECIFICATIONS

CRT Workshop #2

(Location)

(Date)

Sponsored by:
(Sponsors)

For more information call:

86

Developing Test Specifications

Agenda

8:30-9:00	Registration and coffee
9:00-10:00	Agenda Group problem solving
10:00-12:00	Clarifying objectives for test development
12:00-1:00	Lunch
1:00-1:30	Test length
1:30-2:30	Individual work
2:30-2:45	Test plan for review by teacher teams
2:45-3:15	Software for testing
3:15-3:45	Individual work and/or consultation on software
3:45-4:00	Wrap-up

Test Plan Report

Test Title:

Course: **Grade:**

Purpose(s) For Testing:

Instructional Management:

_____ Diagnosis of individual student instructional needs
_____ Grouping of students for instruction
_____ Guidance of individual students -- determining courses of study for students

Entry and Exit Decisions:

_____ Selection of students for special programs
_____ Certification of students -- mastery of objectives, grade promotion, graduation

Programmatic Decisions:

_____ Survey assessment -- look at achievement trends over time to assist in program planning
_____ Program evaluation -- how well an instructional program is working
_____ Accountability -- reporting how the schools are doing to constituent groups

Other:

General coverage:

Inclusion of goals and objectives on the test:

_____ All goals and objectives in this course/grade will be tested
_____ Selected goals or objectives will be tested:

Distribution of questions across goals and objectives:

_____ Each goal area will have the same number of questions regardless of the number of objectives under each goal. (N=) This means that objectives may have different levels of coverage and some objectives may be skipped entirely.
_____ Each goal area will have different numbers of questions depending on the relative importance of the goal, the amount of time spent on the goal during the course and the number of objectives under each goal
_____ Other:

These plans will enable us to report results by:

____ Goal areas
____ Objectives
____ Other:

Other Considerations or Information About the Test to be Developed:

a. Why tests are being developed. For example: how they fit in with the instructional (and other) strategies being used in the district; how they will fit in which the curriculum; why the district is developing their own tests rather than using off-the-shelf tests.

b. How results will be used. Who will use the results. How the results will be reported to support these uses, especially the unit of reporting. For example, "The results will be used to see how our students are doing on the most important objectives determined previously by district teachers. Students, classrooms and buildings will be profiled by these objectives so that instruction can be modified as needed next year."

c. The broad timeline for curriculum and test development and use. For example, the schedule for curriculum work and how it fits in with the schedule for test development. When the tests will be pilot tested and when the tests will be used for real.

d. Other explanations and information that teachers should have. For example, depending on the situation, we have included coding schemes, overall philosophy, constraints such as testing time or formats required, etc.

Test Specifications Summary Table

Sample Format

Goals/Obis-	# Items On Final Tests	# Items By Format M.C.	Essay	Perf.	Other	# Items By*
Goal 1						
Obj 1.1						
Obj 1.2						
Obj 1.3						
.						
. (Add other objectives for Goal 1)						
.						
Total						
Goal 2						
Obj 2.1						
Obj 2.2						
Obj 2.3						
.						
. (Add other objectives for Goal 2)						
.						
Total						
.						
. (Add other goals and objectives for the course)						
.						
Grand Total # Items						
Estimated Tot. Testing Time						
Estimated Tot. # Sittings						

* You can classify the questions you plan on having on the test in any way that is important to you and that others should know about. Examples include cognitive level, difficulty range, topics, mainland versus local situations, cross-reference to other assessments, etc.

Detailed Test Specifications

(A listing of curriculum objectives with detail added so that the intended content coverage on the test can be documented and reviewed)

Sample format

Test Name:

Course: **Grade:**

Goal 1: (list the text of the first goal)

Objective 1.1 (list the text of the first objective)

Format:

Number of items:

Clarifications: (List any clarifications that are needed in order to ensure that everyone has the same interpretation of the objective. This could include such things as cognitive level, lists of acceptable or necessary topics to cover, definitions of terms, relationship to other objectives, sample items, examples of stems, specifications for distractors, etc. There is no fixed list of what to include. This is a matter of judgment.)

Objective 1.2: (list text of second objective in goal 1)

Format:

Number of Items:

Clarification:

(Continue listing goals and objectives until all are completed.)

91

Test Specifications (Blueprint)

Description

The specifications (blueprint) for a test outline exactly what it will look like in terms of content and length. Content covers such things as how each objective will be measured, clarity on the exact content covered by each objective, and various approaches to assessing the content. Length includes total test length and the relative numbers of questions that cover each objective and skill cluster.

The test specifications will be reviewed by teacher teams prior to development of the item pools so that further clarity or exapansion of coverage of objectives can occur prior to developing questions.

Purpose

We are attempting to build curriculum-referenced tests. The purpose of such tests is usually to be able to infer the degree to which students know information, or are able to do various things. From the sample of questions on the test, we want to infer the degree of mastery of a domain of skills. Therefore, in order to be interpretable, the domain from which we will select representative items needs to be clear.

For example, consider the following objectives:

a. "Know that the sum, difference, and product of integers are always integers." What does "know" mean? How will students act when they "know?" Restate the definition? Pick out examples of the principle? State the principle in symbols? Cite this principle as an axiom when doing proofs?

b. "Explains means of preventing, curing or controlling diseases and conditions." What is appropriate for grade 2? Grade 4? Grade 12? Explain in writing or orally? Recall of information only? What would a teacher teach? What would one assess?

c. "Students will differentiate among forms of propaganda in written selections." What forms of propaganda for grade 5? Will students be able to apply titles to types? Is this all they should be able to do?

The point here is that if we are to know what to teach and what to assess we need to have a sufficient degree of clarity. More clarity than is required takes up unnecessary development time. Less clarity leads to confusion and inability to interpret the results. We will spend some time trying to build experience with knowing what is the necessary degree of clarity.

Test specifications also cover the specifics of total test length, amount of testing time required, and the number of items which will cover each objective.

92

Clarity Questions

Could someone who did not develop the curriculum write test questions that would cover exactly the content you want? Using the objectives, would teachers have a common understanding of what to teach? How will you ensure that the results will tell you what to teach and what students know?

The following represent some things to think about as you check your objectives for just the right amount of clarity.

1. Is it clear what level of cognitive level is being assessed? Recall? HOTS?

2. Should the objective be restricted? Expanded?

3. If objectives are repeated across grades, how will assessment and instruction be different? The same?

4. If objectives are repeated across content areas in the same grade, how will assessment and instruction be different? The same?

5. Does your set of objectives cover everything that is important? Are there holes?

93

Assessment Supplement for Teachers

This supplement to CAPTRENDS is designed for easy reproduction and distribution to teachers. PLEASE SHARE IT!

Thinking Skills: Measuring More Than Recall

When you write test items, do you ask students to do more than just recall facts? Increasingly, educators and the public agree that we want students to do far more than regurgitate knowledge; we want them to use their knowledge productively. So it's a little discouraging to learn that in a recent study of over 300 teacher-developed paper and pencil tests (conducted within the Cleveland Public Schools), 90% of the test items measured recall.

Admittedly, getting beyond recall can be tough. How do we define higher level thinking skills and how do we pose questions or write test items to measure those skills once defined? Here are some simple suggestions that can make it easier.

One popular way to categorize higher order thinking skills involves six levels.* After presenting students with new information, we can assess their ability to deal with that information in various ways:

First, we can ask if students can *recall* the information presented.

Second, we can ask if they comprehended or *understood the information.* If they can recount it in their own words, they probably understood it.

Third, we might ask if students can *apply* the information to a new problem situation. If they solve the problem successfully, they can use the information at their disposal.

Fourth, we can ask them to *analyze* or examine components of the information.

Fifth, students might be asked to combine, *synthesize* or assemble the information from two or more sources to draw a conclusion.

And finally, we might have students make some *evaluative judgment* about the information, expressing their opinions.

There are two possible ways to measure students' skill at each level: Teachers can make questions up, or they can rely on questions provided in instructional materials. Let's explore the second option first.

Analyzing Textbook Assessments

Do the textbooks you use include questions that take students beyond recall? What percentage of the questions posed represent each of the levels specified above? The only way to find out is to analyze the study questions posed in the text. Pick a random sample of three or four chapters of a social studies book, for example, and analyze the study questions. Here's an easy way to find the classification of any particular question:

If you can identify:	The question is testing:	Example
What students must *remember*	Recall	What is the electoral college?
What students must *restate* in other words	Comprehension	How does the electoral college work?
What information is to be *used* to solve the problem	Application	Predict what would happen if the electoral college were eliminated.
What is *broken down* into what parts	Analysis	Differentiate the various roles of the electoral college.
What two pieces of information are to be *combined*	Synthesis	How can the electoral college and the popular vote produce different results?
What students are to express an *opinion* about	Evaluation	In your opinion, should the electoral college be retained or abolished? Defend your choice.

To avoid confusion about the level of any question, remember that levels from application upward to analysis and beyond require recall and comprehension as a prerequisite. That is, if the students cannot recall the information and/or do not understand it, she or he will not be able to use it, analyze it, synthesize it or make a considered evaluative judgment. However, remember to understand that each of these higher levels requires some operation beyond just recalling or understanding.

In addition, remember that a key to fair assessment is to be sure that a good match exists between the levels of questioning used for instruction (i.e., in the text or during recitation) and for testing. For example, it would be grossly unfair to ask students merely to recall information during everyday instruction, then present them with a test demanding skills in synthesis and evaluation. We must teach what we test. This leads us to the issue of writing your own test items.

Teacher-Developed Questions

The questions that guide day-to-day recitation in class and that appear on teacher-developed tests and quizzes determine how students will perceive a teacher's expectations. If those questions tap higher order skills, they will give the message that the teacher values more than recall.

But writing such questions from scratch is far more difficult than recognizing them when they occur in a textbook. Right? Not necessarily. Questions that measure thinking skills are relatively easy to write, if we attend to one key part of the question: The verb or action word that describes the problem to the student. Try this simple plan:

If you want to measure:	Start the exercise with these key words:		Examples
Information recall	list describe define label repeat name	fill in identify what when who when	List the parts of speech.
Comprehension	paraphrase explain review match discuss	translate interpret how why	Explain what purpose the verb serves in a sentence.
Application	apply construct draw simulate sketch	employ restructure predict how	Write a sentence that includes a noun, a verb and direct object.
Analysis	classify dissect distinguish differentiate compare	contrast categorize separate breakdown subdivide	Break down this sentence into its components by diagramming it.
Synthesis	combine relate put together	integrate assemble collect	Combine what you know about good sentences and good paragraphs to write an essay on...
Evaluation	judge argue assess appraise decide defend	rate debate evaluate choose should	Evaluate this paragraph. Is it good? Why or why not?

In Summary

If you analyze tests you have developed in the past, you will gain some insight into your question writing tendencies. What level of skills are you measuring? What level do you wish to measure? Try changing the key words in some of the recall questions and watch the level change. But remember, it's not fair (or valid) to teach at the recall level and test at higher levels — or vice versa. Levels of instruction and assessment must match.

*Bloom, B.S. and others (eds.) *Taxonomy of Educational Objectives: Cognitive Domain.* New York: David McKay Co. Inc., 1956.

Lewiston

•A Learner Objective: Given a set of oral directions, the student places a mark over, under or around a picture. } converted into indiv. assessment of Kdg. positional concept

R01P19A01 (a) V0111 I0161

B Learner Objective: The student uses his/her hands to represent positional words when given an oral command (e.g., Simon Says).

untestable

R01P19C01 (d) C Learner Objective: The student places an object in a variety of positions when given oral directions (e.g., place ball on table, etc.). } same as A

V0112 I0162

W Word Recognition (20)

Phonic Analysis

01 Skill: Letters of the Alphabet (-*)

General Objective: Students will identify the letters of the alphabet, both upper case and lower case.

R01W01A01 (c)• A Learner Objective: Given a set of letters, the student identifies the letter named by the teacher.

I0163

R01W01A02 (a) I0164

untestable B Learner Objective: Given a set of letters in random order, the student names the letters.

02 Skill: Consonant Sounds (-*)

General Objective: Students will recognize single consonant sounds.

R01W02A01 (b)• A Learner Objective: Given a list of words with different beginning sounds, the stude reads the letter representing the beginning sound. recognition

I0165

R01W02B01 (c)• B Learner Objective: Given a list of words with different ending sounds, the student reads the letter representing the ending sound. recognition

I0166

03 Skill: Vowel Sounds: Long Vowel Sounds (-*)

General Objective: Students will recognize the long vowel sounds.

R01W03A01 (b) I0167 • A Learner Objective: Given a list of words orally, the student identifies the long vowel words. (R01W03A02 (b) I0168) (combined to make items)

04 Skill: Vowel Sounds: Short Vowel Sounds (-)

General Objective: Students will recognize the short vowel sounds.

R01W04A01 (a) I0169 • A Learner Objective: Given a list of words orally, the student identifies the short vowel words.

05 Skill: Vowel Sounds: Digraphs (.-)

General Objective: Students will recognize vowel digraphs.

R01W05A01 (a) V0113 I0170 • A Learner Objective: Given a picture and an unfinished word that names the picture, the student identifies the missing vowel digraph from a choice of various vowel digraphs.

06 Skill: Vowel Sounds: Dipthongs (.-)

General Objective: Students will recognize vowel dipthongs.

R01W06A01 (c) V0114 I0171 • A Learner Objective: Given a picture and an unfinished word that names the picture, the student identifies the missing vowel dipthong from a choice of various vowel dipthongs.

07 Skill: Consonant Blends and Digraphs (.-)

General Objective: Students will recognize the sound spelled by consonant combinations.

R01W07A01 (b) I0172 • A Learner Objective: Given a set of pictures, the student identifies the consonant letter combinations contained in each word.

R01W07B01 (d) I0173 • B Learner Objective: Given a list of words orally, the student identifies the consonant letter combinations.

untestable • C Learner Objective: Given a list of words, the student writes down the consonant letter combinations contained in each word.

08 Skill: Rule Application: Closed Vowel Short (CVC) (.-)

General Objective: Students will recognize closed vowel short (CVC) words.

A Learner Objective: Given a list of closed vowel short (CVC) words, the student states the closed vowel short (CVC) rule.

09 Skill: Rule Application: Final e Rule (.-)

General Objective: Students will recognize words containing the pattern for the final e rule.

A Learner Objective: Given a list of words containing the pattern for the final e rule, the student states the final e rule.

10 Skill: Rule Application: Hard and Soft Sound of "c" and "g" (.-)

General Objective: Students will recognize the hard and soft "c" and "g" sounds.

R01W10A01 (b) V0116 I0174

• A Learner Objective: Given a group of pictures, the student identifies the hard and the soft sounds for "c" and "g".

Structural Analysis

12 Skill: Compound Words (.-)

General Objective: Students will identify compound words in a given set of words.

R01W12A01 (a) I0175

R01W12A02 (b) I0176

• A Learner Objective: Given a list of words, the student identifies compound words.

13 Skill: Contractions (.-)

R01W13A01 (b) I0177

General Objective: Students will identify contractions in a given set of words.

R01W13A02 (a) I0178

• A Learner Objective: The student matches the given contraction with its definition.

14 Skill: Base Words, Affixes (.-)

• General Objective: Students will identify words that contain affixes in a given set of words.

98

Revised 10/30/87

GRADE 6 AND 8 CBM TEST SPECIFICATIONS
Introduction

Attached are specifications for the Grade 6 and 8 CBM tests revised as the result of meetings with OIS on 10/26 and 27, 1987.. The purpose of the specifications is to outline in more detail what the content of the tests should be. This outline includes a description of the types of questions to be written, the format of questions and the number of questions of each type. For each FPO we:

1. **List and number each PE.** Each PE has been assigned a number. For example, A2 would be the second PE listed under cluster A. These represent a downward extension of Grade 10. Any PEs in grades 6 or 8 that are repeated in grade 10 have the same code (except where noted). New codes were assigned to PEs as new ones emerged.

2. **List any SAT items which might measure each PE.** The reading comprehension, language arts and mathematics portions of the SAT were cross-referenced to PEs. New questions will not have to be written to cover these areas.

3. **Describe any supplementary questions that might be written to cover each PE.** The proposed format for these supplementary items is indicated (multiple-choice, self-rate or teacher rate) as is the proposed number of supplementary questions on the final forms of the tests. (Approximately 1 1/2 times this number of questions will be written initially.)

In preparing the draft test specifications, we have made the following assumptions:

1. SAT, Intermediate 2, Form E, 1982 edition is given in grade 6 and Advanced is given in grade 8.

2. The subtests that are given are reading, math, language arts and writing. Science and social science are not given state-wide.

3. All knowledge-type questions will be four- choice multiple-choice.

4. We will cover only the objectives that are listed. Many skills in the various areas are not included in the PEs. For example math does not cover rounding, place value, inequalities and prime numbers.

5. There will be about 10 data-points per cluster. This can be made up of any combination of multiple-choice, self-rate or teacher-rate items. Some clusters have been combined when the items written to measure them are similar, or when they are short.

6. Grade 6 will have multiple-choice and teacher-rate items; grade 8 will have multiple-choice and self-rate items.

7. When PEs are repeated across grades we assume that they are to be assessed with items of relevance to that grade level, not with "mastery" items.

8. We tried to provide adequate coverage of the PEs, while at the same time keeping the total test to a manageable length.

100

Revised 10/30/87

GRADE 6 CBM
DRAFT TEST SPECIFICATIONS

FPO I: BASIC SKILLS

Cluster A: Oral Communication

A1. Adapts speech to informal and formal situations within the experiences of the student.
A2. Contributes to the completion of a prescribed task through the use of group discussion.
A6. Communicates effectively in conversation with others, in classroom discussions, and in small group interactions.
A7. Gives and responds to oral directions, descriptions, non-verbal messages, and common visual symbols.
A8. Relates experiences, feelings, information, and opinions orally.
A9. Asks questions necessary to gain assistance and/or information.

SAT Items: None

Supplement Type: Teacher Rating, 10 items.

Specifications: We propose that the oral communication rating form for grade 6 be adapted from the grade 10, 8 and 3 measures. A1 and A2 are the same as on grade 8 and 10; A6 is the same as grade 8; and A7, A8, A9 are the same as in grade 3. Note: OIS reviewers would like to produce the revisions of Grade 10 for Grade 6.

Cluster B: Reading

B4. Reads a selection from a variety of materials used by the student describing a situation and its outcome, and determines a probable cause of the outcome.

SAT Items: Reading Comprehension: #6, 34, 36, 60.

Number Items on Supplement: None

B6. Reads and explains simple maps, charts, graphs, tables and illustrations.

SAT Items: See FPO I, G1.

Supplement Type: Multiple-choice, 4 items.

Specifications: This PE overlaps with G1. Since G1 covers charts, graphs and tables, the items in B6 will cover maps and illustrations. Two items will be written on a map and two on an illustration. These items will be of the following types:

1. Find information using symbols.
2. Compare information.
3. Interpret information. This could include what the map or illustration is about, what it might be used for, and what implications could be drawn from it.

B7. Reads an article or selection from a variety of materials used by the student and tells the important details in sequence.

SAT Items: Reading Comprehension: #23.

Supplement Type: Multiple-choice, 1 item.

Specifications: There will be one question written based on a passage used also for another PE. This question will ask about the order in which events occurred. What happened first, what happened last, or what happened before or after something else.

B8. Reads a news article, identifies the main idea, documents it with supporting details, and gives a title or heading appropriate to the article.

SAT Items: There were items on the SAT covering these skills (Reading Comprehension #8, 15, 40, 44), but none of them were based on a newspaper article. The OIS reviewer stated that it might not be necessary to have the items based on a newspaper article. Unless we hear otherwise, we will assume that this is the case and that there will be NO supplementary items.

Number of Items on Supplement: None

Cluster C: Written Communication

C1. Writes letters for various purposes and audiences.
C2. Writes a composition giving information and/or expressing opinions.
C3. Writes a composition to promote ideas using relevant supporting details.
C4. Presents ideas in writing in an orderly manner.
C5. Uses words, sentence patterns, and the conventions of written language appropriately.
C6. Selects and uses writing as a means of expressing feelings and ideas.

SAT Items: C5 is covered by the language arts subtest of the SAT. The OIS Reviewers agreed that the writing sample measures the other PEs.

Number Items on Supplement: None

Cluster D: Math

D3. Solves simple ratio, proportion, and percent problems.
D11. Uses ratios to compare quantities and measurements of objects.

Note: These PEs are combined because simple ratio problems often entail comparing quantities and measurements of objects.

SAT Items: Math Computation: #42, 43, 44 (proportion). Math Applications: #2 (proportion).

Supplement Type: Multiple-choice, 4 items.

Specifications: SAT items did not cover ratio or percent. The two ratio problems should be like those on the Grade 10 CBM. Simple percent problems could include:

1. What percent one kind of object is of a total (e.g., what percent of the students are girls).
2. What percent has been used (e.g., percent of allowance is spent, percent of water is drunk).

D6. Uses whole numbers and commonly used fractions (e.g., 1/4, 1/2) to communicate physical quantities (How many, How much, etc.).

SAT Items: Math Concepts: #13, 19 (number line). Math Applications: #15 (discount), 16 (map scale).

Number Items on Supplement: None

D7. Adds and subtracts whole numbers; multiplies any whole number by a 2-digit number; and divides any whole number by a 1-digit number.

SAT Items: Math Computation: #1, 2, 3, 4, 5 (+); 6, 7, 8, 9 (-); 10, 11, 12, 14, 15, 16 (x); 17, 18, 19, 20, 21, 22, 23 (/). Math Applications: #5 (+); 3 (-); 4, 12 (x); 6 (/).

Number on Supplement: None

D8. Adds and subtracts like-denominator fractions and commonly-used decimals.

SAT Items: Math Computation: #34, 36 (+ fractions); 28, 29 (+ decimals); 31, 32 (- decimals). Math Applications: #9, 11, 38 (+ decimals); 14, 28, 37 (- decimals).

Supplement Type: Multiple-choice, 2 items.

Specifications: Only subtracting fractions was not covered on the SAT. We will write two questions involving subtracting like-denominator fractions with no mixed numbers.

D9. Multiplies and divides decimals.

SAT Items: Math Applications: #8, 10 (x).

Supplement Type: Multiple-choice, 2 items.

Specifications: The SAT covered only multiplication. We will write two that cover division at the sixth grade level.

D10. Does arithmetic mentally.

SAT Items: None

Supplement Type: Teacher Rating, 6 items.

Specifications: We will try out the following scheme to see whether it will work -- To avoid the problem of using the same problems over and over, we will try having a pool of items for the teacher to choose from. The pool will be organized by item difficulty (3 levels) and item type:

1. The teacher reads three to four numbers in a row and the student does the arithmetic mentally, e.g., 3 + 5 - 2 +1. This would be best for items that do not require grouping.
2. The teacher shows the problem to the student and the student computes the answer without pencil or paper. The problems are all numbers; no word problems. These problems can involve some simple grouping symbols and knowledge.
3. The teacher provides problems for the student to estimate the answer mentally. Problems will be of the type one digit x two digit or two digit x two digit.

D12. Adds and subtracts commonly used fractions (mixed and common) with unlike denominators.

SAT Items: Math Computation: #35 (+); 37, 38 (-).

Number Items on Supplement: None

D13. Multiplies and divides mixed and common fractions.

SAT Items: Math Computation: #39, 40, 41 (x). Math Applications: #1, 7 (x).

Supplement Type: Multiple-choice, 2 items.

Specifications: The SAT only covers multiplication. We will write two division problems that are at the sixth grade level. One problem will have common fractions and the other will have mixed numbers.

Cluster E: Measurement

E4. Computes measurement using the four basic operations.

SAT Items: Math Applications: #32, 33, 34, 35, 36, 39.

Number Items on Supplement: None

E5. Estimates measurements.

SAT Items: None

Supplement Type: Multiple-choice, 1 item.

Specifications: As on the grade 3 test, the student will estimate how long, big or heavy something is. The objects will be common for grade 6 students, such as the length of a new pencil, the weight of a chair, etc.

E6. Measures length, capacity, time, temperature, and mass (weight) of objects using standard units.

SAT Items: None

Supplement Type: Multiple-choice, 4 items.

Specifications: These items will entail having a scale of some kind laid against the object to be measured. The student reads the weight, temperature, etc., from the measuring device that is provided. The emphasis will be on capacity, temperature and weight since time and length were covered in E4, above. For each there will be questions covering such things as:

1. Read a value.
2. Compare values.
3. Critical values.

We will have more than one question on each graphic.

Cluster F: Geometry and Cluster G: Graphing

Note: These are combined because of the few number of PEs in each.

F1. Uses correct terminology in describing the properties of geometric figures.

SAT Items: None

Supplement Type: Multiple-choice, 2 items.

Specifications: The intent of this PE was clarified by OIS to be the ability to provide word definitions or descriptions of various properties. For grade 6 appropriate ones would be square, parallel lines or a circle.

F3. Classifies plane and solid geometric figures into various subsets using different specialized properties.

SAT Items: Math Applications: #29 (congruency); 31 (faces).

Supplement Type: Multiple-choice, 2 items.

Specifications: Based on comments from the Grade 10 CBM, the intention of the PE is not such things as naming solids, identifying diameters, or finding parallel or perpendicular lines. Rather, the student should identify which group of objects is alike; the student has to abstract the relevant property, it is not presented in the item. Properties would include right angles, parallel lines, and quadrilaterals.

F4. Identifies, names, and draws various geometric figures.

SAT Items: None

Supplement Type: Multiple-choice, 2 items.

Specifications: Write two questions (one a plane and one a solid) which cover:

1. Identifying the shape named in the stem.
2. Providing the shape and asking the student to select what it is.

G1. Reads, makes and interprets graphs, tables, and commonly-used schedules.

SAT Items: Math Applications: #23, 24, 25 (graph); 26, 27 (table).

Supplement Type: Multiple-choice, 2 items.

Specifications: The two supplementary questions should cover making graphs. The two items should cover a bar and a pictograph. A set of data will be given to the students and they will be asked to identify the graph that best represents the data. These should reflect grade 6 level of difficulty.

SUMMARY OF GRADE 6 CBM ITEM TYPES AND NUMBERS

FPO		KNOWLEDGE ITEMS # TASK	# SUPPLE.	# TEACHER #SELF-RATE	RATES	TOTAL
I	A1, A2, A6, A7, A8, A9	0	0	0	10	10
	B4	4	0	0	0	4
	B6	See FPOI G1	4	0	0	4
	B7	1	1	0	0	2
	B8	4	2	0	0	6
	C1, C2, C3, C4, C5, C6	Writing Sample and L A Subtest	0	0	0	0
	D3/D11	4	4	0	0	8
	D6	4	0	0	0	4
	D7	27	0	0	0	27
	D8	12	2	0	0	14
	D9	2	21	0	0	23
	D10	0	0	0	6	6
	D12	3	0	0	0	3
	D13	5	2	0	0	7
	E4	6	0	0	0	6
	E5	0	1	0	0	1
	E6	0	4	0	0	4
	F1	0	2	0	0	2
	F3	0	2	0	0	2
	F4	0	2	0	0	2
	G1	5	2	0	0	7
	Total	77	49	0	16	142
II	A8	0	0	0	1	1
	A9	0	2	0	0	2
	A10	0	2	0	0	2
	B10,11, 14,15	0	2	0	0	2
	B12	0	0	0	5	5
	B13	0	2	0	0	2
	Total	0	8	0	6	14

FPO		KNOWLEDGE ITEMS # TASK	# SUPPLE	# TEACHER #SELF-RATE	RATES	TOTAL
III	A1	0	3	0	0	3
	A2,A5	0	3	0	0	3
	A3,A4,A7	0	4	0	0	4
	A6	0	2	0	0	2
	A8	0	0	0	4	4
	Total	0	12	0	4	16
IV	A1	0	3	0	0	3
	A2	0	0	0	5	5
	A3	0	0	0	2	2
	A4	0	0	0	1	1
	A5	0	1	0	0	1
	B2	0		0	2	2
	C1Covered by B2 and A2		0	0	0	
	C3	0	5	0	0	5
	C4 Covered by III					
	C5	0	0	0	3	3
	Total	0	9	0	13	22
V	A1	0	1	0	0	1
	A6	0	5	0	0	5
	A7	0	2	0	0	2
	A8	0	2	0	0	2
	B1	0	2	0	0	2
	B2	0	4	0	0	4
	B3	0	4	0	0	4
	B4	0	2	0	0	2
	B6	0	1	0	0	1
	B7	0	1	0	0	1
	B8	0	1	0	0	1
	C1	0	0	0	3	3
	C6, C7, C4	0	0	0	3	3
	C5	0	3	0	0	3
	D1	0	1	0	0	1
	Total	0	29	0	6	35
VI	B6	0	3	0	0	3
	B9, B10, A12	0	2	0	0	2
	D6, D7	0	2	0	0	2
	C4	0	3	0	0	3
	C5	0	2	0	0	2
	C6, C7	0	2	0	0	2
	C8	0	3	0	0	3
	Total	0	15	0	6	21

FPO		KNOWLEDGE ITEMS # TASK	# SUPPLE.	# TEACHER #SELF-RATE	RATES	TOTAL
VII	A5	0	3	0	0	3
	A6	0	3	0	0	3
	A7	0	1	0	0	1
	B5		3	0	0	3
	B6					
	C5					
	C6	0	0	0	8-10	8-10
	C7					
	Total	0	10	0	8-10	18-20
VIII	A1					
	A4	0	0	0	10	10
	A5					
	A6					
	B1/B10	0	0	0	5	5
	B6/B7	0	5	0	0	5
	Total		5		15	20

GRADE 6 CBM

FPO	M.C. # to write	M.C. # on final	Tchr.Rate # to write	Tchr.Rate # on final
I	45	30	16	16
II	12	8	5	5
III	18	12	4	4
IV	13	9	13	13
V	45	29	6	6
VI	22	15	6	6
VII	15	10	10	10
VIII	8	5	15	15
Total	169	112	57	57

CRT MODELS
(Revised May 1979)

OBJECTIVE

Describes the student's task.

ITEM SPECIFICATIONS

STEM

A description of the problem or task given to the student. e.g., a formula, expression, or set of data which the student must use in order to arrive at the correct answer.

RESPONSE ALTERNATIVES

CORRECT RESPONSE

Brief description of the correct response.

DISTRACTORS

A description of the wrong answers. Common error patterns are often used here.

ITEM FORMAT

Multiple choice.

DIRECTIONS AND SAMPLE ITEM

A sample test item consistent with the item specifications listed in the model.

Basic Algebra 1-2
Basic Algebra 3-4

PROGRAM Algebra 1-2 SUBDIVISION II.A.7.

BEHAVIOR LEVEL Application SKILL/CONCEPT Distributive Axiom

OBJECTIVE

The student will identify a correct application of the distributive property of multiplication with respect to addition from a list of mathematical statements.

ITEM SPECIFICATIONS

STEM

Any arithmetic expression of the form $a(b + c) = ab + ac$ with positive integral values less than 10 for a,b,c.

RESPONSE ALTERNATIVES

CORRECT RESPONSE

The mathematical statement that is the correct application of the distributive property. In some instances, "none of the above" will be the correct response.

DISTRACTORS

Other wrong answers will be of the form:

$a(b \times c) = (a \times b)(c)$ See (b) below.

$a(b + c) = (b + c)(a)$ See (c) below.

$a(b \times c) = (c \times b)(a)$ See (d) below.

In some instances "none of the above" will also be a wrong answer.

ITEM FORMAT

Multiple choice.

DIRECTIONS AND SAMPLE ITEM

Which of the following illustrates the distributive property of multiplication over addition?

* A. $2(3 + 5) = 2(3) + 2(5)$

B. $2(3 \times 5) = (2 \times 3)(5)$

C. $2(3 + 5) = (3 + 5)2$

D. $2(3 \times 5) = (3 \times 5)(2)$

E. None of the above

Created June 1979 CRT MODEL

PROGRAM Algebra 1-2 SUBDIVISION III.F.

BEHAVIOR LEVEL Application SKILL/CONCEPT Word Problems (6)

OBJECTIVE

The student will read a word problem and recognize from a list of equations the correct equation used to arrive at the solution.

ITEM SPECIFICATIONS

STEM

Word problem having ticket costs of 60 cents and 25 cents. Number of each kind of ticket is a multiple of ten. Total value of tickets should not exceed $35.00. Solution of the form $60a + 25(t - a) = s$ where a is a multiple of ten and t is less than 100. s is less than 3500.

RESPONSE ALTERNATIVES

CORRECT RESPONSE

The equation that will lead to the solut on. In some instances, "none of the above" will be the correct response.

DISTRACTORS

Other wrong anwers will be of the form:

$60a + 25(a - t) = s$ See (a) below

$60a + 25(a - t) = \frac{s}{100}$ See (c) below

$60a + 25(t - a) = \frac{s}{100}$ See (d) below

In some instances "none of the above" will also be a wrong answer.

ITEM FORMAT

Multiple choice.

DIRECTIONS AND SAMPLE ITEM

Adult tickets to the talent show cost 60 cents each and children's tickets cost 25 cents. Seventy tickets were sold for 28 dollars. How many of each kind of ticket were sold?

If x = the number of adult tickets sold, which of the following equations would lead to the solution?

A. $60x + 25(x - 70) = 2800$

* B. $60x + 25(70 - x) = 2800$

C. $60x + 25(x - 70) = 28$

D. $60x + 25(70 - x) = 28$

E. None of the above

Created June 1979 CRT MODEL

PROGRAM Algebra 1-2 SUBDIVISION X.A.1.b.

BEHAVIOR LEVEL Application SKILL/CONCEPT Solving Quadratic Equations $ax^2 = b$

OBJECTIVE

The student will solve a quadratic equation of the form $ax^2 = b$

ITEM SPECIFICATIONS

STEM

$ax^2 = b$, b is a multiple of a, $0 < b < 22$, $\frac{b}{a}$ is not a perfect square.

RESPONSE ALTERNATIVES

CORRECT RESPONSE

The correct solution of the equation. In some instances, "none of the above" will be the correct response.

DISTRACTORS

Other wrong answers will be of the form:

$\pm \sqrt{b - a}$ See (b) below

$\frac{b}{a}$ See (c) below

$\sqrt{\frac{b}{a}}$ See (d) below

In some instances "none of the above" will also be a wrong answer.

ITEM FORMAT

Multiple choice.

DIRECTIONS AND SAMPLE ITEM

Solve: $2x^2 = 10$

* A. $+\sqrt{5}$ or $-\sqrt{5}$

B. $\pm 2\sqrt{2}$ or $\pm\sqrt{8}$

C. 5

D. $\sqrt{5}$

E. None of the above

Created June 1979 CRT MODEL

PROGRAM Algebra 1-2 SUBDIVISION X.A.3.

BEHAVIOR LEVEL Application SKILL/CONCEPT Quadratic Equations $ax^2 + bx = 0$

OBJECTIVE

The student will solve an equation of the form $ax^2 + bx = 0$

ITEM SPECIFICATIONS

STEM

$ax^2 + bx = 0$; a, b are prime; $2 \leq a, b \leq 7$

RESPONSE ALTERNATIVES

CORRECT RESPONSE

The correct solution of the equation. In some instances, "none of the above" will be the correct response.

DISTRACTORS

Other wrong answers will be of the form:

$-\frac{b}{a}$ See (a) below

$\frac{b}{a}$ See (c) below

0 or $\frac{b}{a}$ See (d) below

In some instances "none of the above" will also be a wrong answer.

ITEM FORMAT

Multiple choice.

DIRECTIONS AND SAMPLE ITEM

Solve: $2x^2 + 3x = 0$

A. $-\frac{3}{2}$

* B. 0 or $-\frac{3}{2}$

C. $\frac{3}{2}$

D. 0 or $\frac{3}{2}$

E. None of the above

Standard: The students will draw, label, and interpret cell models.

Objective: Describe the functions of the following cell parts: nucleus, chromosomes, cell membrane, vacuole, mitochondria, cytoplasm, cell wall, and chlor plasts.

Description of Item Stems:

1. Each test item will require students to select the statement which best describes the function of the cell part named in the stem.

2. The stem for each item will be worded as follows:

   "What is the function of (***) in a cell?

3. The cell part used in place of (***) in each stem will be drawn from the following replacement set: nucleus, cell membrane, vacuole, cell wall, cytoplasm, mitochrondria, chromosomes.

Description of Answer Choices:

1. The correct answer will consist of a short sentence that correctly describes the function of the cell part named in the stem.

2. Each distractor will consist of a short sentence which correctly describes the function of a cell part other than the part named in the stem.

Sample Item:

1. What is a function of the cell wall in a cell?

   a. It controls all cell activities.
   b. It produces energy for all cell activities.
   c. It controls the movement of substances into the cell.
   d. It gives the cell its shape.

Test Specification Exercise

Purpose: Practice in completely specifying how an objective will be measured.

Disclaimer: We do not necessarily recommend that this be done with all objectives. The trick is to know when this level of detail is necessary and when it is not.

Activities:

- Work in district teams.
- (20 min.) Choose one to three objectives that can be measured in multiple-choice format. Have two to three people work on each objective. Develop detailed test specifications using the attached worksheet. (This worksheet combines the formal approaches taken by Utah and Glendale.)
- (20 min.) Review each other's specifications; expand, clarify, discuss.
- Choose a person to report back to the whole group.

Objective Specification Worksheet

Course: **Grade:**

Goal:

Objective:

Number of Questions on Test:

Number of Questions to Write:

Clarification of Objective: (This can include cognitive level, expansion to cover all relevant aspects, reduction to clarify what is taught, relationship to other objectives, other important features, and/or clarification of terms.)

Description of Item Stems (As Needed):

Description of Answer Choices (As Needed):

Other clarifying statements (As Needed):

Sample Item(s) (As Needed):

Test Length and Distribution Among Objectives

A test can only be just so long. Given the amount of time you are willing to devote to testing, there can only be a certain number of questions. This section will provide help on how to determine the length of a test and how to divide it up between the content to be covered. Specifying numbers of test questions is part of your test specifications and will also help in balancing ideal and practical goals for the test.

1. **Determine ideal test length.** How long would the test be if you tested everything at the level of detail that would be ideal? To do this step you need to look at your objectives and decide what you would like to cover on the test, and then decide how many test questions it would take to cover the area adequately.

   a. The first step is to indicate, for each objective, **how it will need to be measured** -- structured (multiple-choice, T/F, matching), essay, short answer, other performance. Most knowledge and some HOTS objectives can be tested in structured format. Oral language, writing, some HOTS, and objectives requiring manipulating objects cannot. These might need to be assessed in a performance test (e.g., writing sample, essay, perform a task) or individually (e.g., oral reading).

      It is, of course, most efficient to use structured format items, especially multiple-choice. Sometimes you can alter the objectives slightly so that they can be written for multiple-choice. For example, "Given a list of words with different beginning sounds, the student reads the letter representing the beginning sound." Change to "Given a list of words with different beginning sounds, the student chooses the letter representing the beginning sound." It is a matter of judgment how much altering can be done and still keep the intent of the objective.

      Some objectives might require other assessment formats such as teacher ratings or self-report of behavior. This can especially happen in the affective domain. For example, "Explains how one's perspective has been broadened through the study of a new language and its culture." "Cooperates with others to attain common goals and objectives."

   b. **For the structured format portion** decide on ideal test length -- the number of questions or testing situations that would let you know how well students have mastered an objective. Things to consider are:

      - You need a minimum of 5-6 questions in order to separately report a skill.
      - You don't need to have the same numbers of questions per objective.

      For example, you might have 35 reading objectives in the third grade. You want to test them all so that you can provide diagnostic feedback to teachers. This requires 5-6 test questions per objective. This would make a total test length of 175-210 questions.

      Another example is that you have 58 grade 6 reading and language arts objectives to assess on the same test. These objectives are broken up into 15 skill areas (goals or clusters). You want to report information to teachers by skill

area, not objective because the test is designed to be more of a survey test than a diagnostic test. You don't want the same numbers of questions for each skill cluster because they have different numbers of objectives. You will make the number of questions for each cluster proportional to the number of objectives. Five clusters will have 8 questions, five will have 10, three will have 12 and two will have 15. This results in 156 questions.

c. For the performance and individual portions decide on the ideal number of exercises. For example, to assess writing ability you might want to have three writing samples -- narrative, social and expository.

2. Estimate how long it will take students to take the ideal test. The typical sitting should be no longer than 15 minutes in grade K, 20 minutes in grade 1, 30 minutes in grade 2 and 45 minutes in grades 3-12. This accommodates attention span and class period length.

   a. For structured items students can answer anywhere from 25 to 90 questions in a sitting depending on their complexity and the grade level of students. For example in grade 6 math you could give 75 computation questions but only 25 word problems. In reading you could give 90 vocabulary items but only 25-30 reading comprehension questions. In general, fewer questions can be given in a sitting when they are complex and when children are younger.

      Given our previous examples, 175-210 grade 3 questions would take about three to five sittings to give; 156 grade 6 questions would take about two to three sittings.

   b. Performance and supply type items will take varying amounts of time. An essay could take 15-30 minutes. Use your best judgment to come up with estimates.

3. Decide if this amount of testing is acceptable and rework question numbers as needed. If testing time is OK, you know how many questions to produce for each objective and you are ready to begin. If not, you will need to modify your plans by:

   a. only testing some of the objectives;
   b. combining objectives;
   c. altering more objectives to be assessed in structured format; or
   d. not reporting results by objective but by skill cluster.

In the grade 3 reading example, you may not be able to report diagnostic information on all objectives. You may need to combine objectives and report at the skill cluster level. (Students screened at this level could then take more detailed tests in those areas in which they are weak.) Or, you may decide that only 20 of the 35 objectives are really crucial and you will only test those.

In the grade 6 reading/language arts example you may decide that the 15 skill clusters can be combined into 12 and that only 130 questions are needed.

Test Specification Exercise

Purpose: Practice in the entire specification process for one course.

Tasks:

1. Look at each objective and indicate:

   a. The required testing format -- multiple choice, matching, T/F, essay, performance, teacher rating, student self-report. Some objectives can be altered slightly to make them suitable for testing in multiple-choice format. (Do this when it is judged that the spirit of the objective is not changed too drastically.)

   b. Whether the objective needs more clarity. Some of these may be easy enough to clarify now. Others may need more work at home.

   c. The ideal number of items or performance situations that would be required to get a good feel for how well students have attained each objective.

2. Estimate the time required to give the ideal test.

3. Modify your plans to accommodate testing time. Modifications may include combining objectives, only generating scores at the cluster level not the objective level, only testing some of the objectives, and/or modifying more objectives so that they are testable in structured format.

4. Next to each objective (or each specification if you have listed specifications under objectives) indicate how many items or situations you need to generate for the item pools. This number will be 1 1/4 to 2 times the number of items needed for the final test forms.

5. Summarize the final numbers on one page. (See Grade 6 CBM for an example.)

If you have to choose between getting greater clarity on some objectives or making it through all the steps above, please try to get through all the steps. It may be hard to estimate the numbers of items needed if an objective is very unclear, but this probably will happen only rarely. Just mark objectives needing a lot of further clarity and proceed with assigning numbers.

Sample Letter To Teacher Review Teams

Dear ,

We appreciate the time you are contributing in order to review the attached blueprints for our curriculum-referenced test in _____. Since the tests will be used to provide us all with information to help us plan instruction, it is very important we obtain feedback from all users so that the tests will cover the information needed.

The attached test blueprint specifies what will be included on the test. Test questions will be generated to match the blueprint. Reviewing the blueprint will help ensure that the test questions generated will adequately cover the information, skills and abilities that we all feel are important.

General information about the rationale, purposes and uses for this test is provided in the attachment, along with detailed information about how each objective will be measured and the number of questions on the test to cover each objective.

Please review these specifications for the following:

1. Does the content proposed for each objective "get at" the spirit of the objective? Would you add any other aspects of the objective that should be assessed? Delete any? Are there any aspects that should be emphasized more than other aspects?

2. Are the number of questions proposed for each objective acceptable? Do the numbers reflect the relative importance of the objectives? What would you change?

3. What other comments, concerns and/or suggestions do you have? Please feel free to comment on any aspect of the testing plan.

Please feel free to write comments directly next to the objectives.

Please provide us your comments by _____. Thank you again for your time.

Sincerely,

Proposed Session #3
Developing Item Pools

The next session is two days in length. The proposed content for the next workshop in the series is:

1. Problem solving and discussion of test specifications.

2. Planning the item pools -- coding, word processing, how to handle graphics, whether to try using an existing bank, etc.

3. Quality control steps for item pools -- technical review, content review, bias review.

4. Writing test questions -- multiple-choice, T/F, matching, essay, performance.

5. Individual work in finding and/or writing questions.

123

Homework For Session #2

Please try to have the following completed for at least one course:

1. Complete your test specifications using the Test Plan Report or another format that contains all the information.

2. Have the specifications reviewed by a relevant teacher team.

3. Revise the specifications.

4. Bring the revised specifications to the next session. These will be used to generate item pools.

HARD COPIES OF TRANSPARENCIES

125

DEVELOPING TEST SPECIFICATIONS

CRT Workshop #2

Developing Test Specifications

Agenda

9:00–10:00 Group Problem Solving

10:00–12:00 Clarifying objectives for test development

12:00–1:00 Lunch

1:00–1:30 Test length

1:30–2:30 Individual work

2:30–2:45 Test plan

2:45–3:15 Software for testing

3:15–3:45 Individual work

3:45–4:00 Wrap–up

127

Test Plan Report
Outline

Background

- o Why tests are being developed
- o How results will be used
- o Timeline for development and use
- o Other explanations

Summary of length and distribution

Test specifications

Test Specifications

Blueprint for content and length.

If we are to infer student performance on a domain of skills from performance on a sample from the domain, the domain needs to be clearly defined so that we can adequately sample it.

129

EXAMPLES OF OBJECTIVES NEEDING MORE CLARITY

The student will read a word problem and recognize form a list of equations the correct one to arrive at a solution.

Measure length, volume, time, temperature and mass using the metric system.

Identify the place value of a digit in a given whole number.

Recognize long vowel sounds.

Clarity

Will people interpret the objective the same way?

- Cognitive level
- Comprehensiveness
- Right amount of detail
- Fit across grades
- Fit across content areas
- Holes

131

Worksheet

Course: Grade:

Goal:

Objective:

Number questions on test:

Number questions to write:

Clarification:

Item Stem:

Answers:

Other:

Sample:

132

Test Length

1. Determine format for each objective
2. Decide on ideal test length
3. Estimate testing time
4. Compromise
5. Summarize

Proposed Session #3
Developing Item Pools

1. Group problem solving
2. Planning item pools
3. Quality control
4. Finding existing items
5. Writing test questions
6. Individual work

134

Homework
Session #2

1. Complete test specifications
2. Specifications reviewed
3. Specifications finalized
4. Bring specifications to #3

135

CURRICULUM REFERENCED TEST DEVELOPMENT

WORKSHOP #3

Developing Item Pools

PRESENTOR'S OUTLINES

Overview of Session and Item Pool Logistics
Presenter's Outline

Purposes: To provide an overview of the two days.
To review progress to date.
To provide general information to participants about considerations in developing item pools.

Format: Lecture/discussion

Materials: Workshop title transparency
Agenda handout and transparency
Sample standard format for multiple-choice questions handout
Item pool logistics transparency
Overhead projector and screen

Presentation Time: One hour

Points To Make:

1. Go over agenda. The purpose of the two days are to prepare for, and begin to assemble an item pool from which questions for the test being developed will come. In order to do this we will cover certain logistical questions, and guidelines for preparing questions of various sorts. We are emphasizing multiple-choice questions because that is what most participants will be using. The entire second day will be devoted to pulling and writing items to match local curriculum objectives.

2. Ask participants: What they brought for today's session; whether they want a different emphasis on any topics; if they have any news to share about what has transpired since last we met.

3. Item pool logistics. There are many little things that can make the development process less than efficient. Some of these are:

   - Item codes and other hidden information. All items should be coded to correspond to the curriculum objective being measured. Other codes that could be added are cognitive level, pointers to graphics or textual passages, and cross-references to test specifications. A sample is:

     IIIA1.10

     where III is the content area, A is the goal, 1 is the objective, and 10 is a pointer to a graphic. The codes could be "hidden" -- arranged so that they may or may not be printed out at will.

     Other hidden information to be attached to each item could include: right answer, scoring procedure, pointers to other items that should not be on the same test, source of item, etc. The advantage to including this information right along with the item is that it makes item review and later item use more efficient. If you print a parallel coded and uncoded

version of your test, teachers not only have the scoring key, but they can also see which items are intended to measure which objectives.

Having these codes will also be very useful if you begin an item bank -- putting each item on a card so that they can be mixed and matched later.

Participants may want to take a moment to develop their coding system.

- Format standards. You should develop a common set of formats for all items so that they all look the same. Format includes rules for punctuation, spacing, capitalization and layout. A sample set is attached. There is no standard in the field; everyone seems to have their own standards.

  Participant may want to take a moment to think about format standards they want to adopt.

- Handling graphics and text. Ask participants about their software capabilities. We recommend using at least a word processor that allows you to have the formatting and characters you need. Science and math tests have to consider special characters and spacing requirements. It's handy to be able to hide text.

  Many available computer software programs will not allow you to produce the graphics you might need. You can handle this by entering text and leaving spaces for the graphics. Be sure to code the graphics and have graphic code pointers in your items so that you know later which graphic goes with which item(s). This is especially important if you begin an item bank. When a graphic goes with more than one item you don't want to have to include it with each item. So, you store graphics separately and have items point to the one(s) you need.

- Teachers can write 10 to 40 items a day; they can pull 30 to 100 items a day. We recommend only writing items if you can't find them elsewhere. Teachers do not necessarily need to write the items in order to feel ownership in the test. They can contribute previous items, review test specifications and review item pools.

- Develop pools with about 1.5 to 2 times as many items in it as you will need on the final test forms. Leave items unnumbered until pilot forms are produced.

Writing Multiple-Choice Questions
Presenter's Outline

Purposes: As an aid in reviewing/revising previously written items.
Some items will have to be written to fill in gaps where previous items could not be found.

Format: Lecture/discussion

Materials: Guidelines for writing multiple-choice questions handout and transparencies
Sample multiple-choice questions for critique handouts
Blank transparencies
Overhead projector and screen

Presentation Time: Two hours

Content of Presentation:

Emphasize again that we are discussing multiple-choice first because many people are interested in this type of item. We will also discuss other item types later.

1. Give test in franzipanics to see which errors they can detect in the items. Go over the answers.

2. The next part of the workshop will be on how to avoid the types of problems encountered in franzipanics. There is a research basis for many of the following recommendations on how to write multiple-choice questions. Researchers have examined the effect of various item features on testing time, test reliability, item discrimination and item difficulty. The recommendations to be covered come from two classic textbooks on the topic and a recent summary of research.

3. Go over item types transparencies and when each should be used (as needed). Emphasize the desirability of having essay and performance items as well as structured items to measure HOTS. Different item types can serve different purposes.

4. Definitions (H p. 3). Go over terms we will use to describe the different parts of a multiple-choice item.

5. General concepts (H, pp.3-4).

   Number of responses to use. Balance test length (the longer the more reliable) against inflated scores due to guessing. Three OK for high ability students if all choices are good. Use four or five if testing a range of abilities. The major consideration is the quality of the distractors. Three is better than four if you can't think of a fourth distractor.

   Item formats. Completion and internal blanks take longer without improving reliability. Complex multiple-choice take long and decrease discrimination, take longer to construct, and can give clues to test takers (e.g., if you know C is not right then you can eliminate all answers that include C). We are not saying that

these formats should never be used, because sometimes it is the best way to assess what you want. Just avoid them if they are not necessary.

5. Stems. We will not cover all the rules; only the ones that seem most important. Cover the rules on the transparencies. Extra comments for some of the rules are:

   3b. Avoid vocabulary that is unnecessarily difficult. Make questions only as complex as they need to be. You don't, for example, want a student to not be able to answer a math problem because he or she can't read the problem. Eliminate extraneous factors that will affect performance.

   3c. Unneeded information increases testing time and decreases reliability.

   5a. Negatives increases the difficulty and increases processing time. Never have double negatives. Sometimes negatives can't be avoided; sometimes you are interested in their ability to find non-examples of something. We're not saying never use negatives, just avoid them when they are not needed.

6. Options. Again we will only cover the more important considerations.

   3c. This can be especially a problem with HOTS items where one can think of scenarios where a "wrong" answer could be "right."

   4a. "None of the above" decreases reliability and item discrimination. Use "None..." with computation items, not with definition items. "I don't know" -- there exists cultural differences in willingness to guess. "All of the above" can increase idfficults and decrease discrimination.

7. Summary of item review steps. Because we've covered a lot of things to remember about items, the summary provides a systematic approach for reviewing items.

8. Go over the sample items on transparencies and critique them as a group.

9. Critique and rewrite sample test questions in the handouts. Depending on the needs of the group these can be done as a large group, in small groups with one person reporting back, or as individuals with volunteers reporting their rewrites. Problems with questions are:

   Question 1 -- wordy, question not clear, typo in choice "c", choice "d" is silly.

   Question 2 -- choices not equal in length nor parallel in amount of information; format problem in choice "d".

   Question 3 -- wordy, capitalize most, grammar clue, not question format, choices mislettered.

   Question 4 -- complex format, typo in stem.

   Question 5 -- not question format, emphasize negative.

   Question 6 -- word clue, more than one right answer, presence of absolutes in some choices.

10. Have participants write items and critique each others; or, participants can critique items that they brought with them.

Writing Items in Other Formats
Presenter's Outline

Purposes: Discuss suggestions for writing items in other formats besides multiple-choice.

Format: Lecture/discussion/group and individual work

Materials: Handout entitled "Principles of Item Writing"
Handout entitled "Federal Way Writing Assessment"
Transparencies for completion, essay, true/false and performance type items
Overhead projector and screen

Presentation Time: One hour

Points to Make:

1. So far we have discussed mainly multiple-choice questions because most individuals are interested in having at least part of their items be in this format. But, multiple-choice cannot be used to assess all possible learner outcomes. This part of the workshop discusses true/false, essay, completion and performance items.

   Some of the rules for writing these types of items are the same as for multiple-choice: use simple wording (unless complex wording is required to assess the objective adequately), avoid negatives, and avoid clues to the answer.

   (Go over the additional rules for each item type and illustrate them with the sample items provided. Many sample items come from a item-writing workshop developed by ETS. The handout comes from materials developed by Mike Hiscox and Evelyn Bryzinski at Northwest Lab.)

2. Completion items are like multiple-choice items without the choices. The advantage is that students have to produce the answer rather than select the answer. We are using completion to include short answers. The items are used to assess knowledge and some thinking skills.

   Go over the rules in the handout and on the transparency. Illustrate the points with the sample items.

3. Longer open-ended answering procedures and essays are good for assessing higher order thinking skills. They should require students to use knowledge in various ways -- compare and contrast, speculate on reasons, provide evidence for a position, describe probable outcomes, critique, analyze, etc. Essays can also be used to assess the ability to organize information and present an argument.

   The first step in writing an essay question is to make the task clear -- what the task is, what the criteria are for judging answers, how long students will have to complete their answer and how many points they will get for their answer.

   The most difficult thing about essays is scoring them consistently. (Remember that our emphasis in this workshop series is more formal assessment in which large groups of students will all be taking the same or parallel items. Therefore,

scoring consistency is very important.) Standardized scoring requires criteria for scoring, anchor papers illustrating scores, trained raters and blind rating. (The Federal Way handout illustrates one scoring system.) Participants will have to decide how to handle grammar, spelling, capitalization, punctuation, etc.

4. True/false are like multiple-choice items that have only two possible answers. Therefore the rules for writing them are very similar to those for multiple-choice. The major extra rules are to be sure that the questions are either always right or always wrong, and that true and false items are the same length.

5. Performance assessment consists of having students perform the actual task on which they will be graded, for example; running, writing, filling out forms, designing an experiment, planning a trip, etc. They are used to look at process as well as outcome, to see whether students can perform real-life tasks and to assess higher-order thinking skills.

   As with essay items, in order to use performance in a standardized assessment, the task and scoring procedures need to be well-defined. The transparencies come from a workshop designed by Rick Stiggins at Northwest Lab. They describe the steps in designing a performance assessment. There is an accompanying booklet called "Watching Children Grow" that can be purchased for workshop participants from _____.

6. Encourage participants to choose formats that will adequately get at what they want to assess in the easiest manner possible. A good combination is multiple-choice following by one or two essays or performance situations.

7. This discussion has not touched on other formats for assessment such as teacher ratings, student self-report of behavior, and portfolios. Some of these might be useful for assessing affective and behavioral goals. The presenter will have to use his or her own judgment on whether to discusses these formats.

Higher Order Thinking Skills
Presenter's Outline

Purposes: To give participant's ideas on how to incorporate higher-order-thinking skills questions into their tests.

Format: Lecture/discussion

Materials: Handouts and transparencies covering two classification schemes, sample items, and HOTS terms
Overhead projector and screen

Presentation Time: One hour

Points To Make:

1. It is important to incorporate HOTS into tests in the proportion that they are emphasized in the classroom because students catch on pretty fast that what is on the test is what is really important. There has been considerable research that shows that most questions on tests in grades K-12 are recall.

   There is some controversy about the ability to assess HOTS in objective format. Some objectives lend themselves very nicely to objective format -- e.g., interpreting graphs, making inferences, math word problems, etc. They have been on achievement tests for years. Others are more problematic, especially in the social sciences, where the objective is how to explore issues that may not have one right answer; or where the goal is to look at process rather than outcome. This might require essay, extended answer or performance tasks.

   In science and math there is generally only one right answer to a problem. Therefore, many science and math problems lend themselves nicely to objective format. (The participants can brainstorm HOTS-type objectives that would and/or would not lend themselves to objective format.)

2. We are providing two taxonomies of HOTS skills to generate some ideas on how to do this. The first is a modification of Bloom's Taxonomy; the second is Robert Ennis' attempt to combine and cover many classification schemes into one comprehensive system. Note that there are other taxonomies, e.g. New Jersey Test of Critical Thinking (Philosophy For Children).

3. The Bloom's Taxonomy approach looks at the level of questioning. Go over the levels. These cut across all subject areas. Go over keywords, question stems and quiz on levels. Advantages are: fast, easy to apply, most teachers are already familiar with it. Disadvantages are: may not cover all the skills we are interested in. For example: affect, creativity, Piaget.

4. The second approach is a skill's approach. The example is Robert Ennis' taxonomy. Briefly go over the taxonomy: affect and cognitive components; the various types of skills; how they apply to various subject areas. (Could use the assessment instrument types transparency to illustrate the various types of skills in the taxonomy. Use the items on the transparencies to give examples of the various types of items. For example, Critical Thinking is illustrated by "cars"

and "safe water." Developmental is illustrated by "path of balls," and creativity is illustrated by "fluency."

One example of trying to write multiple-choice questions to cover more of a skills approach is provided in the handout titled "FPO III". A1-A6 show the objectives to be measured. Underneath each objective is an elaboration of the types of objective-format questions that might be generated to measure each objective. The actual items are attached.

5. The handout titled "Critical Thinking Vocabulary" provides a list of HOTS terms. This is provided for two reasons. First, students need to know what these terms mean as a basis for HOTS. Second, they can be used as vocabulary items on tests or can be incorporated into test questions.

6. Issues in writing HOTS questions (use the items on transparencies to illustrate these points):

   a. For objective format questions you must have only one right answer. Unfortunately, some of the HOTS skills we want to teach deal with how to handle situations that do not necessarily have one right answer.

   b. Also the right answer is a proxy for the thinking skill we are trying to measure. We assume that if the student got the question right, he or she got it right because they applied the thinking skill we wanted them to. This is a validity issue. For example, we have found many other reasons for getting questions right or wrong: guessing, clues, having a good reason why a "wrong" answer is right, and having a good reason why the "right" answer is wrong.

   c. How to disentangle evidence of thinking from cultural and other background influences. General knowledge, assumptions, and what is accepted as right or wrong can differ between cultures.

   d. Previous knowledge can make a HOTS question a recall question. For example, the angle of incidence equals the angle of reflection; what were the causes of the Civil War.

7. What we have found in assessing HOTS in objective format:

   a. Kids can have trouble with HOTS vocabulary

   b. Reading level severely affects performance

   c. Length of passage severely affects performance

   d. Kids can have trouble putting themselves in another's shoes. They have trouble separating their own opintions of things from what another might feel. (E.g., pretend that you are a judge; what would be important to you to know?) They have trouble entertaining other opinions.

Individual Work On Developing Item Pools
Presenter's Outline

Purpose: To start participants on their own item pools.

Format: Individual work

Materials: Test Plans developed by participants as homework from the previous session
Pools of items brought by the presenter for participants to use.

Time: Three to six hours

Procedure:

Participants should begin matching the items brought by themselves and the presenter to their test specifications. It is a good idea to have xeroxing facilities available. If such facilities are not available, we have had participants mark items with yellow stickies; we then xerox the items later for the participants. Participants should mark each item selected with a code that references it to the objective being measured (see the first part of the workshop). They should pick out 1 1/2 times as many items as they will need on the final forms. They should edit at least the first 10 items that they find using the criteria discussed previously. This will give the presenter(s) an opportunity to check on participant knowledge. If participants can't find enough items, then they should write enough items to fill in the gaps. Everyone should have a prototype test by the end of the day.

Homework and Session #4
Presenter's Outline

Purposes: To provide information about the next workshop in the series.
To go over the homework for CRT #3.
To decide when and where CRT #4 will take place.

Format: Lecture/discussion

Materials: Homework and session #4 handout
Overhead projector and screen

Presentation Time: 30 minutes

Points to Make:

1. The content of CRT #4 is default. It can be modified as the participants see fit.

2. The purpose of the homework is to complete the item pools, get them reviewed and finalize them. The pilot tests will be generated from this pool. The next session will go over pilot testing and how to get the tests ready for pilot-testing.

PARTICIPANT HANDOUTS

CURRICULUM REFERENCED TESTING SERIES

WORKSHOP #3 -- DEVELOPING ITEM POOLS

AGENDA

Day 1

8:00 - 9:00	Registration and coffee
9:00 - 9:30	Review of progress to date Overview of the activities for the two days
9:30-10:00	Logistics in developing item pools
10:00-10:30	Writing multiple-choice questions
10:30-10:45	Break
10:45 - 12:00	Writing multiple-choice questions (cont.) Exercises in writing multiple-choice questions
12:00 -1:00	Lunch
1:00 - 2:00	Writing questions in other formats
2:00 - 2:15	Break
2:15 - 3:30	Ensuring that questions cover higher-order-thinking skills
3:30 - 4:00	Preparing for Day 2

Day 2

9:00 - 12:00	Finding/writing test questions
12:00 - 1:00	Lunch
1:00 - 1:30	Planning for session #4
1:30 - ?	Finding/writing test questions

Sample Standard Format For Multiple-Choice Questions

1. When the stem is a question:

   a. each distractor begins with a capital letter;
   b. phrases or other incomplete sentences in a distractor do not have periods; and
   c. complete sentences in distractors have periods.

2. When the distractors complete the stem (i.e., the stem is not a complete question):

   a. each distractor has a period at the end if it completes the stem;
   b. distractors do not begin with capital letters unless the first word is a proper noun.

3. Punctuate and capitalize the following distractors in the following manner regardless of the stem or other distractors:

   a. None of the above
   b. All of the above
   c. The _____ needs no correction.

Example items:

1. What word below is spelled correctly?

   a. Receive
   b. Recieve
   c. Receeve
   d. None of the above

2. John Beluchi is MOST famous for:

   a. how he died.
   b. Saturday Night Live.
   c. his many movies.
   d. None of the above

3. What is a better way to write the sentence:

   Jack, wanted to send a letter to his brother, so he had to buy a stamp.

   a. Jack wanted to send a letter to his brother so he had to buy a stamp.
   b. Jack, wanted to send a letter to his brother so he had to buy a stamp.
   c. Jack wanted to send a letter to his brother, so he had to buy a stamp.
   d. The sentence needs no correction.

GUIDELINES FOR WRITING MULTIPLE CHOICE TEST QUESTIONS

Judith A. Arter
Evaluation and Assessment Program
Northwest Regional Educational Laboratory
101 S.W. Main, Suite 500
Portland, Oregon 97204
(503) 275-9500

This information has been liberally borrowed from:

Ebel, Robert L. *Essentials of Educational Measurement.* Englewood Cliffs, NJ: Prentice-Hall, 1979.

Gronlund, N.E. *Constructing Achievement Tests* Englewood Cliffs, NJ: Prentice-Hall, 1982.

Haladyna, T.M. and Downing, S.M. The validity of a taxonomy of multiple-choice item writing rules. Arizona State University, 2636 W. Montebello Ave., Phoenix, Arizona 85017, 1987.

EXAMINATION IN FRANZIPANICS

1. What is the primary purpose of the cluss in frampaling?

   A. To remove cluss-prangs
   B. To patch tremalls
   C. To loosen cloughs
   D. To repair plumots

2. The fribbled breg will minter best with an

   A. mors
   B. ignu
   C. derst
   D. sortar

3. Why does the sigla frequently overfesk the trelsum?

   A. All siglas are mellious
   B. Siglas are always votial
   C. The trelsum is usually tarious
   D. No trelsa are directly feskable

4. Trassig normally occurs under which one of the following conditions?

   A. When dissels frull
   B. When lusp trasses the vom
   C. When the belgo lisks easily
   D. When the viskal flans, if the viskal is zortil

5. What probable causes are indicated when tristal doss occurs in a compots?

   A. The sabs foped and the doths tinzed
   B. The kredges roted with the rots
   C. The rakogs were not accepted in the sluth
   D. The polats were thonced in the sluth

6. The mintering function of the ignu is most effectively performed in connection with the

   A. arazma tol
   B. fribbled breg
   C. groshing stantol
   D. frallied stantols

FINAL EXAMINATION IN FRANZIPANICS

KEY:

A B C D A B** Notice the overall pattern. Answers should never be permitted to follow an established pattern. Distractors should be ordered in a random fashion. This random order can be achieved by ordering the distractors from short to long or long to short.

1. *A) To remove cluss-prangs

   Answer repeats a key word that appeared in the stem of the question (cluing); answer contains the only hyphenated word; answer is the longest.

2. *B) ignu

   "a" versus "an"; the last word in the stam "an" requires the next word to begin with a vowel. An examinee could select the right answer simply on the basis of grammar--not job knowledge.

3. *C) The trelsum is usually tarious

   Three distractors contain absolutes (A, B, and D). When in doubt, the examinee should select C because it is relative.

4. *D) When the viskal flans, if the viskal is zortil

   The odd distractor, qualifying "if" phrase, and the longest distractor.

5. *A) The sabs foped and the doths tinzed

   The stem requires a plural answer; all other distractors are singular.

6. *B) fribbled breg

   Inter-item cluing--clued to item no. 2.

I. General Information About Multiple-Choice Questions

A. **Definitions**

Stem. The stem poses the problem to the student. It states or implies a specific question.

Correct Response. The correct or best answer.

Distractors. The wrong answers. Distractors should be definitely less correct than the correct response, but plausibly attractive to the uninformed.

Where is the national government of Great Britain located? (Stem)

a. Berlin (Distractor)
b. Birmingham (Distractor)
c. London (Correct response)
d. Paris (Distractor)

B. **Ideal Number of Responses.** There is no evidence that the more responses the better. The important consideration is the quality of the distractors. Three distractors can be as good as more if they are all plausible. Three distractors also enable the students to complete the test faster. Generally, however, if low achieving are being tested, four or five responses are best.

C. **Best Item Formats.** There is some evidence that items stated as questions are more efficient, less confusing and result in better test characteristics than items requiring completions or items that have blanks in the middle.

What is the approximate population of Denmark? (Item stated as question)

The approximate population of Denmark is: (Completion)

If A implies B whenever B implies A, then B and A are said to be _____ of each other. (Blank in middle)

D. **Complex Multiple-Choice.** There is some evidence that you should avoid the use of complex multiple-choice questions such as:

What statement(s) below are true about our present constitution?

I. It was the outgrowth of a previous failure
II. It was drafted in Philadelphia during the summer of 1787.
III. It was submitted by the Congress to the states for adoption.
IV. It was adopted by the required number of states and put into effect in 1789.

a. I only
b. I and III only
c. II and IV only
d. I, II, III, and IV.

D. Other Considerations

1. It is most efficient to base test items on objectives.

2. Sometimes it is helpful to develop "distractor specifications" -- rules for how to generate distractors. These will, for example, describe that the distractors for a multiplication problem will include adding the numbers, subtracting the numbers, etc. The most common mistakes are included.

3. On any multiple-choice test it is a good idea to balance the key -- have the same number of right answers in the a, b, c and d positions.

II. Guidelines for Writing Stems

1. Design each item to measure an important learning outcome:

a. Deal with important and significant ideas; not with incidental details.

Poor:

This question is based on the advertising campaign of Naumkeag Mills to retain the market leadership of Pequot bed linen. What was the competitive position of Pequot products in 1927?

a. Ahead of all competitors among all customers
*b. Strong with institutional buyers but weak with household consumers
c. Second only to Wamsutta among all customers
d. Weak with all groups of consumers

b. Resist the temptation to increase the difficulty of an item by referring to obscure facts.

2. Present a single clearly formulated problem in the stem of the item:

a. A person should be able to know what the question is asking without having to look at the options. (To test this cover up the answers and see if the question is answerable.)

Poor:

A table of specifications:

a. indicates how a test will be used to improve learning.
*b. provides a more balanced sampling of content.
c. arranges the instructional objectives in order of their importance.
d. specifies the method of scoring to be used on a test.

b. One way to avoid the above problem is to avoid using the subject of a sentence as the item stem and its predicate as the the correct response.

Better:

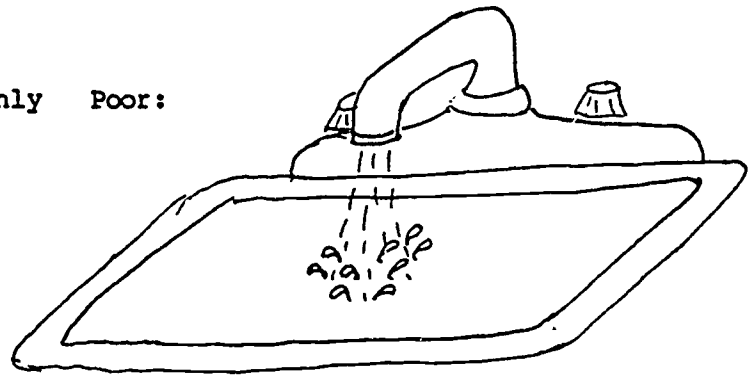
What is the main advantage of using a table of specifications when preparing an achievement test?

a. It reduces the amount of time required.
*b. It improves the sampling of content.
c. It makes the construction of test items easier.
d. It increases the objectivity of the test.

Guideline #2, cont'd.

c. Make sure that the stem asks only one question.

Poor:



What is this?

d. Items which ask for an opinion can have more than one answer depending on one's point of view.

e. If opinions are asked, specify the group or individual whose opinion the examinee is asked to identify.

Poor:

Which event in the following list has been of the greatest importance in American history?

*a. Braddock's defeat
b. Burr's conspiracy
c. The Hayes-Tilden contest
d. The Webster-Hayne debate

Poor:

The second principle of education is that the individual:

a. Gathers knowledge
b. Makes mistakes
c. Responds to situations
*d. Resents domination

3. State the stem of the item in simple, clear language:

a. Avoid complex wording and sentence structure.

b. Avoid vocabulary that is unnecessarily difficult.

Poor:

The paucity of plausible, but incorrect, statements that can be related to a central idea poses a problem when constructing which one of the following types of test items?

a. Short-answer.
b. True-false.
*c. Multiple-choice.
d. Essay.

Better:

The lack of plausible, but incorrect, alternatives will cause the greatest difficulty when constructing what type of test items?

a. Short-answer items
b. True-False items
c. Multiple-choice items
d. Essay items.

c. Avoid information preambles that serve only as window dressing and do not help the examinee understand the question being asked.

Poor:
While ironing her formal., Jane burned her hand accidentally on the hot iron. This was due to a transfer of heat by:

*a. Conduction
b. Radiation
c. Convection
d. Absorption

Poor:
Testing can contribute to the instructional program of the school in many important ways. However, the main function of testing in teaching is:

Better:
What is one function of testing in teaching?

d. If descriptive or qualifying ideas are required place them in different sentences rather than as clauses.

Good:
The term creeping socialism appeared frequently in political discussion in the early 1950s. Which of these is most most often used to illustrate creeping socialism?

*a. Generation and distribution of electric power by the federal government
b. Communist infiltration of labor unions
c. Gradual increase in sales and excise taxes
d. Participation of the United States in international organizations such as the United Nations

4. Put as much of the wording as possible in the stem of the item:

a. Put most repeating words in the stem.

Poor:
Which is the best definition for a vein?

*a. A blood vessel carrying blood going to the heart
b. A blood vessel carrying blue blood
c. A blood vessel carrying impure blood
d. A blood vessel carrying blood away from the heart

b. Reword when options are too long.

Poor:
Instructional objectives are most apt to be useful for test-construction purposes when they are stated in such a way that they show:

a. the course content to be covered during the instructional period.
*b. the kinds of performance students should demonstrate upon reaching the goal.
c. the things the teacher will do to obtain maximum student learning.
d. the types of learning activities to be participated in during the course.

Better:
Instructional objectives are most useful for test-construction purposes when they are stated in terms of:

a. course content.
*b. student performance.
c. teacher behavior.
d. learning activities.

5. State the stem of the item in positive form, whenever possible:

a. The presence of negatives is sometimes confusing.

Poor:
Each question below consists of a word printed in capital letters, followed by five words or phrases lettered a through e. Choose the lettered word or phrase which is most nearly not opposite in meaning to the word in capital letters.

EXAGGERATION:

a. Slight misunderstanding
*b. Silence
c. Accurate representation
d. Truth
e. Understatement

c. Emphasize negative wording whenever it is used in the stem of an item.

Poor:
Which one of the following is not a desirable practice when preparing multiple-choice items?

a. Starting the stem in positive form.
b. Using a stem that could function as a short-answer item.
c. Underlining certain words in the stem for emphasis.
*d. Shortening the stem by lengthening the alternatives.

Better:
Which of the following is NOT desirable to do when preparing multiple-choice questions?

a. Stating the stem in positive form.
b. Using a stem that could function as a short-answer item.
c. Underlining certain words in the stem for emphasis.
d. Shortening the stem by lengthening the alternatives.

6. Use the question format for items. Avoid using completion format and items with blanks in the middle. (See Part I for examples of these.)

7. Avoid using the same questions or problems in a test that were used during instruction.

8. Avoid having the stem or choices in one item give away the answer to another item (item-item effects).

III. Guidelines for Writing Alternative Responses and Avoiding Clues to the Right Answer

1. Use clear, simple and short wording in the alternative responses:

a. Keep the responses as short as possible. This can often be done by rewording the stem.

Poor:
What is monogamy?

a. Refusal to marry
b. Marriage of one woman to more than one husband
c. Marriage of one man to more than one wife
*d. Marriage of one man to only one wife

Better:
What is a marriage called in which one woman marries one man?

a. Unicameral
b. Dualism
c. Monotheism
d. Monogamy

b. Avoid unnecessarily complex vocabulary or sentence structure.

2. All of the distractors should be plausible:

a. Each distractor should relate to the question.

Poor:
The chief difference between the surface features of Europe and North America is that
a. The area of Europe is larger.
b. Europe extends more to the south.
c. The Volga River is longer than the Missouri-Mississippi.
*d. The greater highlands and plains of Europe extend in an east-west direction.

b. Avoid obscure distractors.

Poor:
A chaotic condition is
a. Asymptotic
*b. Confused
c. Gauche
d. Permutable

c. Avoid obvious tricks.

Poor:
Horace Greeley is known for his
a. Advice to young men not to go West
b. Discovery of anesthetics
*c. Editorship of the New York Tribune
d. Humorous anecdotes

d. A distractor which is absurd or highly improbable will contribute nothing to the effectiveness of the item. However, the use of humor, per se, is not prohibited.

Poor:
Which of the following has helped most to increase the average length of human life?
a. Fast driving
b. Avoidance of overeating
c. Wider use of vitamins
*d. Wider use of inoculations

3. There whould only be one right answer among the options:

a. The alternatives should represent a set of distinct options. The differences between the options should be obvious.

Poor:
Meat can be preserved in brine due to the fact that:
a. Salt is a bacterial poison.
*b. Bacteria cannot withstand the osmotic action of the brine.
c. Salt alters the chemical composition of the food.
d. Brine protects the meat from contact with air.

b. Avoid opinions (see stems).

c. Think through all distractors carefully to make sure that none of them could ever be true.

4. Avoid, or use sparingly, "I don't know," "All of the above," and "None of the above."

   a. Don't only include them in items when they are the correct response.

   b. Only use them as the correct answer when all the alternatives are clearly wrong.

Good:
Which word is misspelled?
- a. Contrary
- *b. Tendancy
- c. Extreme
- d. Variable
- e. None of these

Poor:
What does the term growth mean?
- *a. Maturation
- b. learning
- c. Development
- d. All of these
- e. None of these

5. Avoid structural clues to the right answer:

   a. The responses should be similar in grammatical structure, in length, and in language complexity.

Poor:
Lack of attention to learning outcomes during test preparation:
- a. will lower the technical quality of the items.
- b. will make the construction of test items more difficult.
- c. will result in the greater use of essay questions.
- *d. may result in a test that is less relevant to the instructional program.

Poor:
The recall of factual information can be measured best with a:
- a. matching item.
- b. multiple-choice item.
- *c. short-answer item.
- d. essay question.

Better:
What is the best question type to use for the recall of factual information?
- a. Matching items
- b. Multiple-choice items
- c. Short-answer items
- d. Essay questions.

b. State all alternatives in parallel form.

Poor:

Why should negative terms be avoided in the stem of a multiple-choice item?

*a. They may be overlooked.
b. The stem tends to be longer.
c. The construction of alternatives is more difficult.
d. The scoring is more difficult.

Better:

Why should negative terms be avoided in the stem of a multiple-choice item?

*a. They may be overlooked.
b. They tend to increase the length of the stem.
c. They make the construction of alternative more difficult.
d. They may increase the difficulty of the scoring.

c. Whenever the alternatives form a quantitative scale, they should normally be arranged in order of magnitude.

Good:

What is the approximate population of Denmark?

a. 2 million
b. 4 million
c. 7 million
d. 15 million

6. Avoid verbal clues to the right answer:

a. Do not repeat key words from the item stem, or their synonyms in in the correct answer. (These can, however, sometimes be put in the wrong answer.)

Poor:

Which one of the following would you consult first to locate research articles on achievement testing?

a. Journal of Educational Psychology
b. Journal of Educational Measurement
c. Journal of Consulting Psychology
*d. Review of Educational Research

b. Avoid making the correct response more or less general or inclusive than any distractor.

Poor:

Which one of the following types of test items measure learning outcomes at the recall level?

*a. Supply-type items.
b. Selection-type items.
c. Matching items.
d. Multiple-choice items.

Poor:

History tells us that all nations have enjoyed participation in:

a. Gymnastics
b. Football
*c. Physical training of some sort
d. Baseball

c. Avoid different qualifiers among the right and wrong options (e.g., "always", "never", and "only" in the distractors and and "sometimes", "often" and "unless" in the right answer).

Poor:
Achievement tests help students improve their learning by:
a. encouraging them all to study hard.
*b. informing them of the progress.
c. giving them all a feeling of success.
d. preventing any of them from neglecting their assignments.

d. Avoid words with unfavorable connotations among the wrong answers and words with positive connotations in the right answer.

7. Avoid conceptual clues to the right answer:

a. Avoid "convergence." This occurs when distractors are obtained by varying two dimensions of the right answer. The right answer can sometimes be the only one which fits into both dimensions.

1. a. before breakfast
b. on a full stomach
c. with meals
d. before going to bed

2. a. 7,000
b. 6,900
c. 6,810
d. 6,800

IV. Tactics and Sources for Obtaining Good Distracters

1. Define the class of things to which all the alternative answers must belong.

2. Think of things that have some association with terms used in the question.

3. If the items call for a quantitative answer, make the responses distinctly different points along the same scale.

In many situations the precise value is less important than knowledge of a general level of relationship.

Good:
How did (A) the estimated amount of petroleum discovered in new fields in 1953 compare with (B) the amount extracted from producing fields in the same year?
a. A was practically zero.
b. A was about half of B.
c. A just about equaled B.
*d. A was greater than B.

4. Phrase the question so that it calls for a "yes" or "no" answer plus an explanation.

Good:
Has the average size of farms in the United States tended to increase in recent years? Why?

a. Yes, because as the soil loses its natural fertility more land must be cultivated to maintain the same output
*b. Yes, because the use of farm machiery has made large farms more efficient than small farms
c. No, because the difficulty in securing farm labor has forced many farmers to limit their operations
d. No, because large family farms tend to be subdivided to provide smaller farms for the children

5. Use various combinations of two elements as the alternatives.

Good:
What was the general policy of the Eisenhower administration during 1953 with respect to government expenditures and taxes?

a. Reduction of both expenditures and taxes
*b. Reduction of expenditures, no change in taxes
c. Reduction in taxes, no change in expenditures
d. No change in either expenditures or taxes

6. Include true statements which do not correctly answer the stem question.

Good:
What is the principle advantage of a battery of lead storage cells over a battery of dry cells for automobile starting and lighting?

a. The storage cell furnishes direct current.
b. The voltage of the storage cell is higher.
*c. The current from the storage cell is stronger.
d. The initial cost of the storage cell is less.

7. Familiar expressions and phrases that have been used in common parlance may seem attractive to students whose knowledge is merely superficial – e.g., "good" sounding words, stereotyped phrasing and scientific sounding answers.

Good:
Which of these has effected the greatest change in domestic plants and animals?

a. The influence of the environment on heredity
b. Organic evolution
c. Selective breeding
d. Survival of the fittest

8. Use common misconceptions and errors in techniques

9. If alternatives still remain elusive, consider using a different approach in the item stem.

Principles of Item Writing[1]

	Yes	No	Unsure
1. Have I used items measuring important parts of the curriculum?	✓		
2. Have I avoided using items that are presented in an ambiguous fashion?	✓		
3. Have I followed standard rules of punctuation and grammar in constructing items?	✓		
4. Have I constructed only items that have right or clearly best answers?	✓		
5. Have I kept the reading difficulty of test items low in relation to the group being tested?	✓		
6. Have I constructed test items from statements taken verbatim from instructional materials (for example, textbooks)?		✓	
7. If any items are based on an opinion or authority, have I stated whose opinion or what authority?	✓		
8. Do items offer clues for answering other items in the test?		✓	
9. Do students learn things from items that help them answer other items in the test?		✓	
10. Do any of the items contain irrelevant cues?		✓	
11. Have I made any items overly difficult by requiring unnecessarily exact or difficult operations?		✓	
12. Do any of my items have words such as "always," "never," "none," or "all" in them?		✓	
13. Have I included any "trick" items in the test?		✓	
14. Have I checked the items with other teachers or item writers to try and eliminate ambiguity, technical errors, and other errors in item writing?	✓		
15. Do any of the items try to test more than a single idea?		✓	

[1]Prepared by Ronald Hambleton and Daniel Eignor.

"✓" indicates the correct response.

	Yes	No	Unsure
16. Have I restricted the number of item formats in the test?	✓		
17. Were the most "valid" item formats used in the test?	✓		
18. Have I grouped items presented in the same format together?	✓		
19. Do the correct answers follow essentially a random pattern?	✓		

Multiple Choice Items

1. Is each item designed to measure an important objective?	✓		
2. Does the item stem clearly define a problem?	✓		
3. Have I included as much of the item in the stem as possible?	✓		
4. Have I put any irrelevant material in the item stem?		✓	
5. Have I included any grammatical cues in the item stem?		✓	
6. Have I kept to a minimum the number of negatively stated item stems?	✓		
7. If the negative is used in an item stem, have I clearly emphasized it?	✓		
8. Is there one correct or clearly best answer?	✓		
9. Have I avoided the use of answers such as "all of the above" and "none of the above"?	✓		
10. Have I made sure that all answers are gramatically consistent with the item stem and parallel in form?	✓		
11. Have I avoided stating the correct answer in more detail?	✓		
12. Have I made sure that all distractors represent plausible alternatives to examinees who do not possess the skill measured by the test item?	✓		
13. Have I avoided including two answers that mean the same, such that both can be rejected?	✓		

	Yes	No	Unsure
14. Have I avoided the use of modifiers like "sometimes" and "usually" in the alternatives?	✓		
15. Have I made sure to use important sounding words in the distractors as well as the correct answer?	✓		
16. Are all answers of the same length and complexity?	✓		
17. Have I made the answers as homogeneous as possible?	✓		
18. Have I varied the length of the correct answer, thereby eliminating length as a potential clue?	✓		
19. Have I listed answers on separate lines, beneath each other?	✓		
20. Have I used letters in front of the answers?	✓		
21. Have I used new material for the students in formulating problems to measure understanding or ability to apply principles?	✓		

Matching Items

1. Are the entries in the two sets homogeneous in content?	✓		
2. Are there more answers than premises?	✓		
3. Is each answer a plausible alternative for each premise?	✓		
4. Is the length of the set too long (greater than 8-10 premises)?		✓	
5. Are the entries in the sets arranged in some logical order?	✓		
6. Have I indicated whether an answer can be used more than once?	✓		
7. Do my directions specify the basis on which the match is to be made?	✓		
8. Have I made sure that the matching exercise is on one page?	✓		
9. Have I used headings for the premise and answer choices?	✓		

	Yes	No	Unsure
10. Have I made sure that the information couldn't be better obtained using another format, such as multiple choice?	✓		

True-False Items

	Yes	No	Unsure
1. Would another item format be more appropriate?		✓	
2. Is the item definitely true or false?	✓		
3. Does each item contain a single important idea?	✓		
4. Is the item short?	✓		
5. Is simple language used?	✓		
6. Have I made sure that one part of the item isn't true while another part of the item is false?	✓		
7. Does an insignificant word or phrase influence the truth or falsity of an item?		✓	
8. Have I avoided using negative statements?	✓		
9. Have I avoided using vague words such as "seldom" and "frequently"?	✓		
10. Have I avoided use of words that give clues to the correct answer, for example, "always," "never," "usually," and "may"?	✓		
11. Have I made sure my true statements are no longer in length than my false statements?	✓		
12. Are there approximately an equal number of true and false statements in the test?	✓		

Completion Items

	Yes	No	Unsure
1. Would another item format be more appropriate?		✓	
2. Is the item written so that a single brief answer is possible?	✓		
3. Have I omitted unimportant words?	✓		

	Yes	No	Unsure
4. Have I left too many blanks?		✓	
5. Are the blanks near the end of the item?	✓		
6. Have I avoided the use of specific determiners, such as "a" and "an," and singular and plural verbs?	✓		
7. Have I made sure the length of my answer blank is equal from question to question?	✓		
8. If the problem requires a numerical answer, have I indicated the units I want the answer stated in?	✓		
9. If the problem requires a written answer, do students know how spelling errors will be scored?	✓		
10. Is each problem written so clearly that there is a single correct answer?	✓		

Essay Items

1. Are essay questions only being used to measure higher-order objectives?	✓		
2. Are the questions closely matched to the objectives they were written to measure?	✓		
3. Does each question present a clear task to the student?	✓		
4. Is there sufficient time for answering questions?	✓		
5. Are students aware of the time limits?	✓		
6. Do students know the points for each question in the test?	✓		
7. Have I used new and/or interesting material in my essay questions?	✓		
8. Have I tried to start the questions with words or phrases such as "Compare," "Contrast," "Give the reason for," "Give original examples of," "Explain how," "Predict what would happen if," "Criticize," etc.?	✓		
9. Have I written a set of directions for the essays?	✓		

	Yes	No	Unsure
10. If several essays are used, have I used a range of complexity and difficulty in the questions?	✓		
11. Have I prepared an "ideal" answer to each essay question before administering the test?	✓		
12. Have I allowed the students a choice of questions only in those instances when students aren't going to be compared?	✓		

Sample Multiple-Choice Questions For Critique

1. Some occupations require that you pass certain tests before you can practice them. Henry, a resident of Honolulu, wants more information about what is required to be able to work in one of these occupations. Which of the following will help him the most?

   a. Doctor's Guide to Setting Up Practice
   B. Licensed Occupations in Hawaii
   c. Preparing for the SAT
   d. World Book Encyclopedia

2. What type of training and certification is provided by a community college?

   a. Four year academic degrees
   b. Two year programs which lead to associate degrees in many fields such as nursing, electronic repair and child care.
   c. Professional training
   d. union apprenticeships

3. Jennifer's dad is a plumber. She sees she has some of the same ability he had and she would like a job like his, but she doesn't want to be a plumber. The occupation most like a plumber is an:

   a. carpet cleaner.
   b. electrician.
      legal secretary.
   c. physical therapist.

4. Which of tehse occupations works with data rather than things or persons?

      A. Accountants
      B. Construction Workers
      C. Social Workers
      D. Plumbers

   a. A only
   b. A, B and C
   c. A and D only
   d. A, B, C and D

5. We have laws for all the following reasons except to:

   a. protect people.
   b. protect the environment.
   c. keep people from criticizing the government.
   d. maintain order.

6. What are lobbyists and special interest groups likely to do?

   a. Advertise their products
   b. Get bills favorable to their interests passed
   c. Always present both sides of to an issue to voters
   d. See that streets and roads are kept attractive and clean

Sample Rewrites

1. Some occupations require that you pass certain tests. If you want to learn about what is required to be able to work in one of these occupations which of the following references will give you the MOST complete information?

   a. Doctor's Guide to Setting Up Practice
   b. Licensed Occupations in Hawaii
   c. Preparing for the SAT
   d. Occupational Outlook Handbook

2. What type of training and certification is provided by a community college?

   a. Four year academic degrees which lead to community service jobs such as social work
   b. Two year programs which lead to associate degrees in many fields such as nursing, electronic repari, and child care.
   c. Professional training and degrees for careers which require education after college.
   d. Union apprenticeships such as plumbing, electrical and teamsters

3. Which occupation below would be MOST like a plumber?

   a. Carpet cleaner
   b. Electrician
   c. Legal secretary
   d. Physical therapist

4. Which of the following occupations works primarily with data?

   a. Accountants
   b. Construction Workers
   c. Social Workers
   d. Plumbers

5. Which of the following is NOT a reason that we have laws?

   a. To protect people.
   b. To protect the environment.
   c. To keep people from criticizing the government.
   d. To maintain order.

6. In what kinds of activities are lobbyists and special interest groups LEAST LIKELY to be involved?

   a. Advertising their concerns to the public
   b. Getting laws passed
   c. Presenting both sides of an issue to voters
   d. Talking to lawmakers about their concerns

Measuring Thinking Skills in the Classroom

Revised Edition

Richard J. Stiggins
Evelyn Rubel
Edys Quellmalz

Northwest Regional Educational Laboratory

Produced in cooperation with the NEA Mastery in Learning Project

nea PROFESSIONAL LIBRARY
National Education Association
Washington, D.C.

Table 1
SUMMARY OF THINKING SKILLS

Level	Definition	Relation to Bloom Taxonomy
Recall	Most tasks require that students recognize or remember key facts, definitions, concepts, rules, and principles. Recall questions require students to repeat verbatim or to paraphrase given information. To recall information, students need most often to rehearse or practice it, and then to associate it with other, related concepts. The Bloom taxonomy levels of knowledge and comprehension are subsumed here, since verbatim repetition and translation into the student's own words represent acceptable evidence of learning and understanding.	Recall Comprehension
Analysis	In this operation, students divide a whole into component elements. Generally the part/whole relationships and the cause/effect relationships that characterize knowledge within subject domains are essential components of more complex tasks. The components can be the distinctive characteristics of objects or ideas, or the basic actions of procedures or events. This definition of analysis is the same as that in the Bloom taxonomy.	Analysis
Comparison	These tasks require students to recognize or explain similarities and differences. Simple comparisons require attention to one or a few very obvious attributes or component processes, while complex comparisons require identification of and differentiation among many attributes or component actions. This category relates to some of the skills in the Bloom level of analysis. The separate comparison category emphasizes the distinct information processing required when students go beyond breaking the whole into parts in order to compare similarities and differences.	Analysis
Inference	Both deductive and inductive reasoning fall in this category. In deductive tasks, students are given a generalization and are required to recognize or explain the evidence that relates to it. Applications of rules and "if-then" relationships require inference. In inductive tasks, students are given the evidence or details and are required to come up with the generalization. Hypothesizing, predicting, concluding, and synthesizing all require students to relate and integrate information. Inductive and deductive reasoning relate to the Bloom levels of application and synthesis. Application of a rule is one kind of deductive reasoning; synthesis, putting parts together to form a generalization, occurs in both inductive and deductive reasoning.	Application Synthesis
Evaluation	These tasks require students to judge quality, credibility, worth, or practicality. Generally we expect students to use established criteria and explain how these criteria are or are not met. The criteria might be established rules of evidence, logic, or shared values. Bloom's levels of synthesis and evaluation are involved in this category. To evaluate, students must assemble and explain the interrelationship of evidence and reasons in support of their conclusion (synthesis). Explanation of criteria for reaching a conclusion is unique to evaluative reasoning.	Synthesis Evaluation

ASSESSMENT PLANNING CHART

GRADE LEVEL Elementary SUBJECT English TOPIC Poetry

	ORAL	TEST	PERFORMANCE
RECALL	Let's see if we can tell what poetry is.	A haiku has _____ lines and _____ syllables.	Compare some work of local poets. Choose the one you like best. As you know, artists are often not "recognized" until after their deaths so they often live in poverty. You serve on a board which could help this poet you have chosen. How would you go about convincing the board to give funds to this particular poet? Prepare your materials for convincing. Try them out on the class.
ANALYSIS	Let's see how many different kinds of poetry we can list.	Tell the purpose of an acrostic in poetry and write a brief example.	
COMPARISON	Let's compare several kinds of poetry as to rhyming technique, rhythm, structure, etc. Shall we first compare a limerick and a couplet?	Put an H if a haiku, a C if a couplet, and a D if a diamente. ___a. Has 17 syllables ___c. Has 7 lines ___b. Has 2 lines ___d. Has 3 lines.	
INFERENCE	Have several poems ready for use on the overhead, written on the board, or prepared as handouts. Students choose a poem and imagine how it might sound written as another form. A haiku as a couplet, etc. Put examples on the board.	What kind of poem is likely to be written by a Japanese person? a. Limerick c. Haiku b. Couplet d. Cinquain.	
EVALUATION	Do you think it's important to encourage the writing of poetry? Why or why not?	Some poetry is very structured. Do you prefer it to the less structured? Why or why not? Cite two examples of the kind of poetry you prefer.	

ASSESSMENT PLANNING CHART

GRADE LEVEL Junior High SUBJECT Social Studies TOPIC Electoral College

	ORAL	TEST	PERFORMANCE
RECALL	What is the electoral college?	As a member of the electoral college, you must vote: a. According to your own judgment b. As your constituency voted c. As the party tells you d. Only if you wish to do so.	Assume you're a U.S. senator. Propose a constitutional amendment that would make the popular vote the sole criterion for electing a president. Your amendment would do away with the electoral college. Prepare a speech to Congress defending your amendment. Be sure to— a. Analyze all elements of the issue, b. Compare elections with and without the college, c. Show how the voters are likely to react, and d. State and defend your values. (Or conduct a simulated debate on the Senate floor.)
ANALYSIS	How does the electoral college work?	Analyze the steps in the presidential election process, showing where the electoral college comes into play.	
COMPARISON	How do the social conditions that existed when the electoral college was formed differ from conditions now?	What is meant by the election theme "one person, one vote," and how does that relate to the electoral college?	
INFERENCE	If you were a presidential candidate elected by popular vote, could you still lose the election? How?	In which state is the electorate likely to oppose the use of the electoral college: a. California b. Illinois c. Etc.	
EVALUATION	Should the electoral college be abolished? Why or why not?	Which of the following is the best reason for maintaining the electoral college? a. Tradition. b. Fairness to large states. c. Etc.	

ASSESSMENT PLANNING CHART

GRADE LEVEL — Junior High SUBJECT Science TOPIC Energy

	ORAL	TEST	PERFORMANCE
RECALL	Today we're beginning a unit on energy. Let's develop a definition for the word energy.		You are planning to build a new house. Compare the energy sources. Decide which you'll use in your house. Explain how you'll use the sources you've chosen. Also tell why you chose that source for that particular job. You may draw the plans showing the various places of energy use if you wish. If you do this with a good explanation key, then you need not write an explanation.
ANALYSIS		What are the three effects of mining and burning coal?	
COMPARISON	Let's compare the costs of using the following energy sources: nuclear, solar, coal, water, wood, and wind.		
INFERENCE	If we were suddenly cut off from the supply of petroleum from the Middle East, what would be some short-term problems? What would be some long-term solutions to those problems?		
EVALUATION		Write a short paper telling which energy source is the most important for the United States. Give at least three reasons why it's the most important.	

STUDY STEP 4: ADDING VARIETY TO THE QUESTIONS

Now let's move from the formula chart to the generation of charts with a greater variety of questions.

The first key to expanding the range of questions you can pose is to focus on the trigger or action verb used to describe the problem to the students. Start with these and add some of your own if you can:

If you want to measure:	Use these key words in the exercise:		Illustration
Recall	define identify label list name	repeat what when who	List the names of the main characters in the story.
Analysis	subdivide break down separate	categorize sort	Break the story down into different parts.
Comparison	compare contrast	differentiate distinguish	Compare the themes of these two stories.
Inference	deduce predict infer speculate	anticipate what if . . . apply conclude	How might we make this character more believable?
Evaluation	evaluate judge assess appraise defend	argue recommend debate why critique	Evaluate this story. Is it well written? Why or why not?

The second key to expanding the range of questions you can pose is to plug these action words into a growing list of generic questions. Again, consider these and add some of your own if you can:

Recall

- Define the word _____.
- What is a _____?
- Label the following _____.
- Identify the _____ in this _____.
- Who did _____?

Analysis

- What are the basic elements (ingredients) in a _____?
- What is/are the function(s) of _____?
- Inventory the parts of _____.
- Categorize the _____ of _____.
- Sort the _____.
- Analyze the following _____.

Comparison

- Compare the _____ before and after.
- Contrast the _____ to the _____.
- Differentiate between _____ and _____.

Inference

- Hypothesize what will happen if _____.
- Predict what would be true if _____.
- Conclude what the result will be if _____.
- What if _____ had happened instead?
- What does this information suggest?
- Given this situation (problem), what should you do?
- What rule applies in this case?

Evaluation

- What do you believe about _____?
- Judge what would be the best way to solve the problem of _____. Why did you decide that?
- Evaluate whether you would _____ or _____ in this situation. Why?
- Decide if _____ was worth it. Explain.

Use these lists of action verbs and questions to generate a complete chart on another topic of relevance to you.

FINAL STEP: A PROGRESS CHECK

In the space provided next to each exercise, enter the letter that represents the thinking skill category reflected in the item (see Appendix B for the answers):

R=Recall A=Analysis C=Comparison I=Inference E=Evaluation

_____ 1. What are three functions of the liver?

_____ 2. Let's brainstorm what would happen if the sun did not come up tomorrow.

_____ 3. Define the word *mitasis*.

_____ 4. Which of the following menus is the best? Why?

_____ 5. Which menu provides more complete protein?

_____ 6. Should the use of computers be abolished in the classroom? Why or why not?

_____ 7. Who is the author of *Where the Sidewalk Ends*?

_____ 8. If we mix these chemicals together, what do you suppose will happen?

_____ 9. Look at the chart showing the number of meals Americans have eaten away from home over the last three years. How have eating habits changed?

_____ 10. What are three purposes of an unmanned space flight to Jupiter?

_____ 11. What are the functions of our eyelashes?

_____ 12. Which do you think will have greater impact on your life, the invention of the computer or our ability to travel in space? Why?

_____ 13. If you were going outside and it was snowing quite hard, which of the following would you need from your closet?

____ a. Your umbrella

____ b. Your lightweight jacket

____ c. Your warm boots

____ d. Your sandals.

_____ 14. You hate rain, but know it is necessary. What are three purposes it serves?

_____ 15. In the Northwest it rains and snows a lot. Which is more vital to create the supply of water necessary for summer use?

_____ 16. What are some jobs a migrant worker might perform in getting a crop of lettuce to market?

_____ 17. Haiku is a form of _____.

_____ 18. Look at these three paintings. Which makes the most use of vivid colors?

_____ 19. Suppose we had not dropped the bombs on Hiroshima and Nagasaki. How else might we have defeated the Japanese?

_____ 20. Which is a better snack for you, a fresh peach or a dish of frozen peach yogurt? Why?

ASSESSMENT PLANNING CHART

GRADE LEVEL _____ SUBJECT _____ TOPIC _____

	ORAL	TEST	PERFORMANCE
RECALL			
ANALYSIS			
COMPARISON			
INFERENCE			
EVALUATION			

Homework For Session #3

Please complete the following tasks before the next workshop session:

1. Complete the item pool -- find/write items; enter items; get graphics designed; paste up pools; proof pools for typos, correct graphics, complete and accurate codes.

2. Have item pools reviewed by teachers. Provide a coded version of the pool for teachers to review. They should also get a copy of the test specifications. A prototype cover letter is attached which describes to the reviewers what they should do.

3. Summarize teacher comments, revise items and provide feedback to reviewers on what action was taken in response to their comments.

4. Finalize the pools. Make sure that you still have about 1.5 times as many items as you will need for the final forms. These will be pilot tested. Check this by objective. If you have more than 1.5 times the items (by objective) you may need to weed some out. If you have fewer, you may need to write some extras.

5. Bring your finalized item pools (coded version) to the next session.

188

Prototype Cover Letter For Item Pool Reviewers

Dear _____,

Thank you for assisting us to develop our curriculum-referenced test in _____. Your time will help ensure that we have a quality product that covers the information and skills that we all feel are important.

Please review each test question for the following things. Please be as specific as possible in your comments. For example, if you feel a question does not measure an objective, please indicate why (or write a new question). This will enable us to make revisions that adequately respond to your concerns. You may attach a copy of the questions with comments and revisions written in.

1. Does the question measure the objective that is intended for it to measure? Each question is cross-referenced to the attached curriculum objectives using the following coding scheme:

   (describe your coding scheme)

2. Do the test questions measuring each objective represent a good range of coverage for the objective? Is there anything essential that is left out?

3. Is there any unnecessary complexity in the question? Is there only one right answer?

4. Are the questions of the proper level of difficulty for the students? If not, specifically what should be different?

5. Are there any questions that you feel are inappropriate because of controversial content, bias toward any ethnic group, etc?

6. Do you have any other comments on the tests, procedures or purposes for the tests?

Thank you for your time!

Proposed Content For Session #4

The following topics will be covered:

1. Preparing the item pool for pilot testing -- sample items, directions to students, administration instructions for teachers, covers, etc.

2. Testing logistics -- spiralling tests, numbering tests, answer sheets, distributing and collecting materials, color-coding, etc.

3. Sampling -- how many students to test, how to get a representative sample.

4. Scoring tests -- quality control on answer sheets, reporting formats.

5. Statistics -- what statistics to run on your items and tests and how to use these to improve the items.

190

Federal Way Public Schools

31455 28th Avenue South Federal Way, WA 98003 941-0100 or 927-7420

April 17, 1984

TO: All Junior High Principals and Counselors
All Ninth Grade English Teachers

FROM: Ted Hagen, Director of Testing & Evaluation

RE: Writing Competency Test

The Writing Competency Test will be administered to all 9th. grade students the week of April 23-27. The test takes two full class periods (see instructions to teachers) and must be given on two consecutive days. i.e., Monday, Tuesday, or Tuesday, Wednesday, etc.

Students should be told that this is a "real" test and that it will become part of their permanent record, and, that passing it is a graduation requirement. The test will be scored in May, 1984, and students will be given their scores before schools is out for the summer.

Enclosed are some materials to explain the scoring criteria to you. The tests will be scored by readers who have received special training.

When all testing is done and make-ups are completed, alphabetize the final drafts of the test by school and return to Barbara McCallum by Tuesday, May 1, 1984.

TJH:bmc

Copy: Ted Gartner

FEDERAL WAY SCHOOL DISTRICT

GUIDELINES FOR EVALUATING COMPOSITION, GRADES 1-12

During a workshop in February 1981, twenty-four elementary and secondary teachers developed a set of guidelines f the evaluation of composition. The guidelines were piloted in the spring and the rating scale was revised in June 1981. Junior high and high school English teachers also piloted a writing competency test which they would like the District to adopt. Teachers in the workshops expressed a need for developing a strong district writing program with a systematic approach in assessing student composition.

The categories and criteria of the composition rating scale are intended for formal evaluation which may take place once or twice per quarter. Teachers are not expected to use all the categories in assessing every piece of student writing. They should select one or two categories for instructional emphasis and evaluation in each assignment. Teachers are encouraged to change the point system of the rating scale to suit the purpose of the assignment. The rating scale is more suitable for evaluating expository writing than narrative or creative writing. The category "Originality" is not included in the proposed competency test rating scale, but it is important in classroom instruction and evaluation.

The rating scale provides indicators for arriving at scores one, three, and five. In-between scores of two and four on the scale will be determined by the judgment of the reader/teacher. The reader/teacher is to score each category from one to five. One is the lowest possible score in each category; three is the middle or average score; and five is the highest possible score. Each raw score should be multiplied by the weighting factor of that category to determine the points earned. For example, a three (3) in Organization would earn nine (9) points (i.e., 3 x 3). Using the weighting factors for the competency test, the highest possible score is 100 points; the lowest is 20.

The evaluation of a student's written composition should be based upon expectations appropriate to the grade level and the nature of the assignment. The categories in the rating scale are applicable in grades 1-12. Each paper should be evaluated on the basis of what is actually written, without considering the expectations from the particular student. The criteria alone should be the basis for evaluation.

FEDERAL WAY PUBLIC SCHOOLS
WRITING COMPETENCY TEST

Instructions to Teachers*

The writing competency test has been designed to meet the District requirement that all students demonstrate minimum writing competencies before graduating from high school.

The test will consist of one writing sample which will take two days to complete. Students will spend the first d. ' developing their ideas and drafting them in a rough copy. That copy will be turned in at the end of the test period on the first day and returned to students at the beginning of the test period on the second day. Students will spend that second day editing their work and producing a final copy of the test. At the end of the second testing period, teachers will collect the test papers, affix student labels, and return them to the bulding test coordinator in the envelopes provided. If a student is absent, make-up sessions must be scheduled. The writing samples will be rated by trained readers and results will be available to both students and teachers.

SPECIFIC INSTRUCTIONS

Day 1: Testing time is one class period.

Materials:

1. Writing test instruction sheets for each student.
2. Pencils.
3. Paper.
4. Dictionaries.
5. Thesaurus.
6. Student name labels.

Procedures:

1. Have students read instructions silently as you read them aloud.
2. Remind students that THEY MAY NOT RECEIVE ASSISTANCE from their teacher or classmates.
3. Collect all papers at the end of the test period.
4. Remind students to bring a pen with black or blue ink for the next test session.

Day 2: Testing time is one class period.

Materials:

1. Paper.
2. Pens.
3. Dictionaries.

Procedures:

1. Have students read instructions silently as you read them aloud.

2. Remind students that THEY MAY NOT RECEIVE ASSISTANCE from their teacher or classmates.
3. Remind students to write final copy using pen with black or blue ink.
4. Collect test papers as students finish, affix student labels to upper right-hand corner of the first page of the paper. Papers should be in alphabetical order, place in the envelope provided, and returned to your building test coordinator.
5. If a paper is more than one page in length, be sure the student's name is written on each additional page and that pages are stapled together in the upper left-hand corner.
6. Return all copies of the writing test instructions (student copies) to your building test coordinator.

*Developed by Bellevue School District

Revised: April 1984

FEDERAL WAY PUBLIC SCHOOLS
WRITING COMPETENCY TEST

Instructions to Students*

To the student: (Read silently as your teacher reads aloud.) This test asks you for a sample of your best writing. This writing sample will be rated according to the following list of writing skills.

1. The writer communicates a purpose.
2. The writer provides enough relevant information so that the reader can make sense of the topic.
3. The writer organizes material to develop a topic.
4. The writer uses language appropriate to the reader and the situation.
5. The writer constructs sentences grammatically and applies standard usage.
6. The writer applies basic rules for capitalization and punctuation.
7. The writer spells common words correctly.
8. The writer produces a neat and legible final copy.

GENERAL INSTRUCTIONS

°You will be writing a paper that contains an introductory paragraph, developing paragraph(s), and a concluding paragraph.

°You will have two class periods to write your paper.

°Use the first class period to get ideas, to plan, and to draft your writing. Use your own paper.

°Use the second class period to write your final copy.

°Use a dictionary or thesaurus if you wish, but do not ask the teacher or your classmates for help.

°Give your paper an appropriate title.

°Use blue or black ink for the final copy. Use your own paper. Write on only one side of the page. Do not write in the upper right-hand corner (space measuring 1½" x 4") of the page.

TOPIC: If I Could Do It Over

Everyone at some time has done something that they wished they had done differently. Choose one thing you have done and write a paper in which you explain what you would like to have done differently and why.

*Developed by Bellevue School District
Revised: April 1984

FEDERAL WAY PUBLIC SCHOOLS
WRITING COMPETENCY TEST

Instructions to Readers

For the purpose of clarity in use of the Federal Way Rating Scale, the following comments should be incorporated by the readers:

I. Purpose:

A. Carries main idea or purpose throughout the paper.

B. Follows directions for assessment:
1. appropriate length
2. writes on assigned topic
3. puts a title on paper

II. Information

A. Uses information that is redundant and/or repeated.

B. Uses logical arguments.

III. Organization

A. Uses smooth and effective transitions.

IV. Punctuation

A. Uses correct syllabication when dividing words at the end of a line.

V. Usage

A. Writes numbers correctly.

B. Does not use abbreviations.

VI. Language Variety and Precision

A. Omits word parts.

If the student does not finish the final copy, it is rated on the completed portion only, not the rough draft. Most likely, and incomplete paper would result in failure.

Revised: April 1984

FEDERAL WAY PUBLIC SCHOOLS

Rating Scale for Writing competency Test

WRITING VALUE SCALE

Papers will be scored using the following weighted categories:

1. (10) Purpose/Central Idea
2. (15) Information
3. (15) Organization
4. (15) Language Variety & Precision
5. (10) Sentence Structure
6. (10) Conventional Usage
7. (10) Spelling
8. (5) Capitalization
9. (5) Punctuation
10. (5) Neatness & Legibility
11. * Originality

The number listed to the left of each category is the maximum number of points.

1. Purpose/Central Idea

   5 - States or reveals purpose with exceptionally clear central idea.

   3 - Has a recognizable central idea which may be too general, too specific or too vague.

   1 - Has no recognizable central idea.

2. Information

   5 - Uses sufficient relevant information to support and develop the central idea.
   - Avoids repetition of information.

   3 - Uses some relevant and accurate information to support and develop the central idea.

   1 - Lacks supporting information.
   - Gives irrelevant or inaccurate information.

3. Organization

   5 - Logical sequencing of ideas.
   - Divides topic effectively into paragraphs.
   - Effective movement from beginning to end (introduction - body - conclusion).

   3 - Understandable sequencing of ideas.
   - Usually divides the topic effectively into paragraphs.
   - Adequate movement from beginning to end.

   1 - Faulty or inadequate sequencing of ideas.
   - Does not use paragraphing successfully.
   - Inadequate or faulty movement from beginning to end.

4. Language Variety and Precision

- Uses a combination of general/specific and abstract/concrete language.
5 - Uses words precisely.
- Uses a variety of words.

- Generally uses appropriate language.
3 - May use clichés, slang, or be too wordy, but stays with the central idea.
- Demonstrates an adequate vocabulary.

- Uses inappropriate language.
1 - Demonstrates an inadequate vocabulary.
- Repeats words unnecessarily.

5. Sentence Structure

5 - Structures all sentences clearly and correctly.
- Uses an effective variety of sentence structure.

- Generally structures sentences clearly.
3 - May contain a few fragments and/or run-on errors.
- Usually uses a variety of sentence structures.

- Uses sentence structures that obscure meaning.
1 - Makes excessive run-on sentence and sentence fragment errors.
- Lacks variety of sentence structure.

6. Conventional Usage*

- Uses conventional subject/verb agreement and pronoun/antecedent agreement.
5 - Uses appropriate verb tense consistently.
- Uses adjectives and adverbs (modifiers) appropriately.
- Employs usage appropriate to the topic.

- Usually uses conventional subject/verb agreement and pronoun/antecedent agreement.
3 - Usually uses verb tense consistently.
- Usually uses adjectives and adverbs (modifiers) appropriately.
- Avoids most errors in usage.

- Makes error in basic subject/verb agreement and pronoun antecedent agreement.
- Uses inconsistent verb tenses.
1 - Misuses adjectives and adverbs (modifiers).
- Makes usage errors that interfere with meaning.
- Consistently fails to observe usage conventions.

*Confusion in usage of homonyms is included in this category.

7. Spelling

5 - Spells all words correctly.

3 - Makes some spelling errors, usually in writing difficult words, but also may violate some basic spelling rules.

1 - Makes so many spelling errors that the errors interfere with readability.

8. Capitalization

5 - Observes rules of capitalization.

3 - Usually uses punctuation marks correctly.

1 - Seldom observes rules of capitalization.

9. Punctuation

5 - Uses punctuation marks correctly.

3 - Usually uses punctuation marks correctly.

1 - Seldom uses punctuation marks correctly.

10. Neatness and Legibility

- Arranges the paper neatly on the page.
5 - Writes legibly.

- Produces a paper that shows an attempt at neatness.
3 - Uses handwriting that does not interfere with readability.

- Produces a paper that lacks neatness.
1 - Writes illegibly.

11. Originality

5 - Achieves originality of ideas or approach.

3 - Routine, predictable ideas or approach.

1 - Lacks creativity or imagination in ideas or approach.

Revised: April 1984

The Fourth International Conference on
Critical Thinking & Educational Reform

June 21, 1985

GOALS FOR A CRITICAL-THINKING/REASONING CURRICULUM[1]

Robert H. Ennis
Illinois Critical Thinking Project
University of Illinois, U. C.
1310 South Sixth Street
Champaign, IL 61820

WORKING DEFINITION: Critical thinking is reasonable reflective thinking that is focused on deciding what to believe or do.[2]

Critical thinking so defined involves both dispositions and abilities:

A. DISPOSITIONS:

1. Seek a clear statement of the thesis or question
2. Seek reasons
3. Try to be well-informed
4. Use credible sources and mention them
5. Take into account the total situation
6. Try to remain relevant to the main point
7. Keep in mind the original and/or basic concern
8. Look for alternatives
9. Be openminded
   a. Consider seriously other points of view than one's own ("dialogical thinking")
   b. Reason from premises with which one disagrees--without letting the disagreement interfere with one's reasoning ("suppositional thinking")
   c. Withhold judgment when the evidence and reasons are insufficient
10. Take a position (and change a position when the evidence and reasons are sufficient to do so
11. Seek as much precision as the subject permits
12. Deal in an orderly manner with the parts of a complex whole
13. Be sensitive to the feelings, level of knowledge, and degree of sophistication of others[3]

B. ABILITIES: (Classified under these categories: Elementary Clarification, Basic Support, Inference, Advanced Clarification, and Strategy and Tactics)

Elementary Clarification

1. Focusing on a question
   a. Identifying or formulating a question
   b. Identifying or formulating criteria for judging possible answers
   c. Keeping the situation in mind

2. Analyzing Arguments
   a. Identifying conclusions
   b. Identifying stated reasons
   c. Identifying unstated reasons
   d. Seeing similarities and differences
   e. Identifying and handling irrelevance
   f. Seeing the structure of an argument
   g. Summarizing

3. Asking and answering questions of clarification and/or challenge, for example:
   a. Why?
   b. What is your main point?
   c. What do you mean by "_____"?
   d. What would be an example?
   e. What would not be an example (though close to being one)?
   f. How does that apply to this case (describe case, which might well appear to be a counterexample)?
   g. What difference does it make?
   h. What are the facts?
   i. Is this what you are saying: _____?
   j. Would you say some more about that?

Basic Support

4. Judging the credibility of a source; criteria:
   a. Expertise
   b. Lack of conflict of interest
   c. Agreement among sources
   d. Reputation
   e. Use of established procedures
   f. Known risk to reptuation
   g. Ability to give reasons
   h. Careful habits

5. Observing and judging observation reports; criteria:
   a. Minimal inferring involved
   b. Short time interval between observation and report
   c. Report by observer, rather than someone else (i.e., not hearsay)
   d. Records are generally desirable. If report is based on a record, it is generally best that:

1) The record was close in time to the observation
2) The record was made by the observer
3) The record was made by the reporter
4) The statement was believed by the reporter, either because of a prior belief in its correctness or because of a belief that the observer was habitually correct

e. Corroboration
f. Possibility of corroboration
g. Conditions of good access
h. Competent employment of technology, if technology is useful
i. Satisfaction by observer (and reporter, if a different person) of credibility criteria (#4 above)

Inference

6. Deducing, and judging deductions

a. Class logic - Euler circles
b. Conditional logic
c. Interpretation of statements

1) Double negation
2) Necessary and sufficient conditions
3) Other logical words: "only", "if and only if", "or", "some", "unless", "not", "not both", etc.

7. Inducing, and judging inductions

a. Generalizing

1) Typicality of data: limitation of coverage
2) Sampling
3) Tables and graphs

b. Inferring explanatory conclusions and hypotheses

1) Types of explanatory conclusions and hypotheses

a) Causal claims
b) Claims about the beliefs and attitudes of people
c) Interpretations of authors' intended meanings
d) Historical claims that certain things happened
e) Reported definitions
f) Claims that something is an unstated reason or unstated conclusion

2) Investigating

a) Designing experiments, including planning to control variables
b) Seeking evidence and counterevidence
c) Seeking other possible explanations

3) Criteria: Given reasonable assumptions,

a) The proposed conclusion would explain the evidence (essential)

b) The proposed conclusion is consistent with known facts (essential)
c) Competitive alternative conclusions are inconsistent with known facts (essential)
d) The proposed conclusion seems plausible (desirable)

8. Making and judging value judgments

a. Background facts
b. Consequences
c. Prima facie application of acceptable principles
d. Considering alternatives
e. Balancing, weighing, and deciding

Advanced Clarification

9. Defining terms, and judging definitions: three dimensions:

a. Form

1) Synonym
2) Classification
3) Range
4) Equivalent expression
5) Operational
6) Example - nonexample

b. Definitional strategy

1) Acts

a) Report a meaning ("reported" definition)
b) Stipulate a meaning ("stipulative" definition)
c) Express a position on an issue ("positional", including "programmatic" and "persuasive" definition)

2) Identifying and handling equivocation

a) Attention to the context
b) Possible types of response:

i) "The definition is just wrong" (the simplest response)
ii) Reduction to absurdity: "According to that definition, there is an outlandish result"
iii) Considering alternative interpretations: "On this interpretation, there is this problem; on that interpretation, there is that problem
iv) Establishing that there are two meanings of key term, and a shift in meaning from one to the other

c. Content

10. Identifying assumptions

a. Unstated reasons
b. Needed assumptions: argument reconstruction

Strategy and Tactics

11. Deciding on an Action

    a. Define the problem
    b. Select criteria to judge possible solutions
    c. Formulate alternative solutions
    d. Tentatively decide what to do
    e. Review, taking into account the total situation, and decide
    f. Monitor the implementation

12. Interacting with Others

    a. Employing and reacting to "fallacy" labels (including)

1)	Circularity	12)	Conversion
2)	Appeal to authority	13)	Begging the question
3)	Bandwagon	14)	Either-or
4)	Glittering term	15)	Vagueness
5)	Namecalling	16)	Equivocation
6)	Slippery slope	17)	Straw person
7)	Post hoc	18)	Appeal to tradition
8)	Non sequitur	19)	Argument from analogy
9)	Ad hominem	20)	Hypothetical question
10)	Affirming the consequent	21)	Oversimplification
11)	Denying the antecedent	22)	Irrelevance

    b. Logical strategies
    c. Rhetorical strategies
    d. Presenting a position, oral or written (argumentation)

        1) Aiming at a particular audience and keeping it in mind
        2) Organizing (common type: main point, clarification, reasons, alternatives, attempt to rebut prospective challenges, summary--including repeat of main point)

Notes

1. This is only an overall content outline. It does not incorporate suggestions for level, sequence, repetition in greater depth, emphasis, or infusion in subject matter area (which might be either exclusive or overlapping).

2. Elaboration of the ideas in this set of proposed goals may be found in my "Rational Thinking and Educational Practice" in Jonas F. Soltis (ed.), Philosophy and Education (Eightieth Yearbook of the National Society for the Study of Education, Part I), Chicago: NSSE, 1981; also my "A Conception of Rational Thinking" in Jerrold Coombs (ed.), Philosophy of Education 1979, Bloomington, IL: Philosophy of Education Society, 1980. A note on terminology: the term: "rational thinking", as used in these articles, is what I mean here by "critical-thinking/reasoning". In deference to popular usage and theoretical considerations as well, I have abandoned the more narrow appraisal-only sense of "critical thinking" that I earlier advocated.

3. Item 13 under "Dispositions" is not strictly speaking a critical thinking disposition. Rather it is a social disposition that is desirable for a critical thinker to have.

FPO III Content Description

A1. Identifies, clarifies, and states a problem and develops criteria for examining alternatives in solving the problem.

1.	Identify central issue or problem	Airbags, Pemba, Gorge, Cartoon
2.	Restate problem or paraphrase	
3.	Compare similarities, develop analogies	
4.	Develop criteria for judging	Drugs
5.	Design an experiment	Grades, Detergent, Lightbulbs

A2. Gathers information from various sources and analyzes and organizes the information to facilitate the formulation of alternatives.

1.	What other information is needed	Pemba, Airbags, Detergent, Lightbulbs
2.	What information is relevant/irrelevant	Gorge, Airbags
3.	Where to get information	Pemba
4.	Fact v. opinion/biased v. unbiased sources	Airbags
5.	Checking the consistency of information	Drugs
6.	Identify assumptions	Airbags, Detergent
7.	Appraising observations	Gorge
8.	Which two values are in conflict	

A3. Formulates hypotheses about a problem based on available information.

1.	Develop hypothesis	Detergent
2.	Formulate questions that lead to deeper understanding	

A4. Applies the criteria established to select an alternative.
A5. Evaluates the alternative selected for its effectiveness.

1.	Which proposed solution might be best	Pemba
2.	Applying criteria	Airbags

A6. Draws conclusions or generalizations based on the alternatives or hypotheses and related information.

1.	Inferences and deductions	Gorge
2.	Logical syllogisms	Baboons
3.	Recognizing adequacy of data	
4.	What information supports a conclusion	Grades, Gorge
5.	Cause and effect	
6.	Interpreting the results of an experiment	Detergent

A7. Validates and reports the conclusions and modifications, if any.

1.	What to do to validate choice	Pemba, Airbags
2.	Probable consequences	
3.	How to report results	Detergent
4.	Expanding results	Detergent

The following tells about a fictional court case. Assume that you are a judge in the case as you answer questions ___ to ___.

A Tanzanian district court must decide whether to let a mining operation on Pemba Island go ahead as planned. The mining operation would change the shape of the entire coast of the island. The mine has been held up for over a year by environmental groups who feel it would cause a lot of damage to the special ecology of the island.

Supporters of the mine say that there are no animals on the island that are not also on the mainland or on nearby Zanzibar Island. So, any destruction of the ecology would not cause any extinctions. In addition, the mine's backers say that hundreds of jobs will be created. They also say that the operation will bring millions of dollars of foreign money into Tanzania.

Opponents of the mine argue that no definite studies have ever been made of the ecology of Pemba Island. Therefore, there can be no assurance that environmental changes will not occur. They also feel that the creation of a few hundred jobs will have a small effect on the economy and that extra money will mostly benefit the mining companies and not the people in general.

Under Tanzanian Law, nothing may be undertaken that will cause damage to the environment unless the environment can recover or unless the project is needed to protect the people from harm.

! VIII .01 !

As a judge you need to gather information to decide what to do. What is the most important issue to decide?

a. Whether the economic benefits of the project outweigh the damage to the environment.
b. The extent of the damage to the island and the extent to which it will recover.
c. Where money from the operation of the mine will go.
d. How many of each kind of animal are in places other than Pemba Island.

! III A1.01.1a1, #b !

Which of the following pieces of information would be MOST helpful in reaching a decision on whether or not to allow the mining operation to go ahead?

a. There are only three species of monkeys on the island compared with over 130 on the mainland.
b. People living near the island are in favor of the mine.
c. The distance of the island from the mainland means that the island animal population cannot get to the mainland.
d. Opponents of the mine get money from a number of international groups.

! III A2.01.1a1, #c !

Which of the following would be the BEST source of information about the potential effects of the mine on island ecology?

a. A study published by an agency of the United Nations on the ecology of offshore islands.
b. A visit to the island by you, as judge, in the company of lawyers for both sides.
c. The report by a team of biologists hired by the mining company to study the environmental impact of the project.
d. Testimony by a French zoologist who has been living on the island to study its animal population.

! III A2.01.1a2, #d !

In order to avoid a long court battle, you, as judge, suggest to both sides that they settle their differences out of court. Which of the following solutions is MOST likely to be acceptable to both sides?

a. The mining company agrees to restore the island to its former condition after removal of the minerals.
b. The mining company agrees to donate money to help the environment in other parts of Tanzania to make up for damage to Pemba Island.
c. The environmental groups agree to have a representative of their group help direct the mining company so that damage to the island will be small.
d. The environmental groups agree to allow the mining if the mining company will return some of their profits to the Tanzanian people.

! III A4.01.1a1, #a !

Assume that the two sides come to an agreement out of court. Under the law, you must approve the agreement before it can take effect. The agreement is designed to permit mining and avoid environmental damage. Which of the following criteria would you consider to be the most important that the agreement satisfy?

a. Both sides must show that they entered into the agreement without influence from outside sources.
b. The experts for both sides must verify that, in their opinions, the agreement will accomplish its stated aims.
c. A way must be set up to show that the agreement is being kept.
d. Meeting the conditions of the agreement must not be so expensive that it will be impossibile to keep.

! III A7.01.1a1, #c !

Use the following information to answer questions ___ to ___.

You are a representative in the U.S. Congress. You must vote whether air-bags should be required in new cars. Upon collision air-bags inflate automatically from under the dashborad to protect the driver and passengers from going through the windshield. You have recently heard the following information on air-bags:

a. There is a court case going on right now about whether the federal government can require air-bags if states don't pass seatbelt laws.
b. Safety experts say the issue should not be air-bags versus seatbelts but that both should be used.
c. Auto-makers don't feel its fair to make them have to put air-bags in cars just because people don't want to buckle up.
d. Insurance companies favor air-bags because people don't use seat belts.
e. State laws on seatbelts must include strict enforcement of the law and a $25 fine for those not buckling up.

! vIII .02

!

You must vote on whether air-bags should be mandatory in new cars. What is the central issue you must decide?

a. The safety of air-bags.
b. The relative safety of air-bags and seatbelts.
c. Whether it is in the best interests of the public to require air-bags in new cars.
d. Can the government require citizens to use air-bags and seatbelts if they don't want to.

! III A1.02.1a1, #c !

In their testimony, the safety experts make some statements which <u>assume</u> that other things, not stated, are true. These assumptions include all the following <u>EXCEPT</u>:

a. Air-bags increase safety.
b. People will use seatbelts if they are available.
c. Air-bags will work as intended.
d. Car-makers will not put air-bags in cars unless forced to.

! III A1.02.1a2, #b !

What other information is LEAST useful for you to know to decide if air-bags should be mandatory in automobiles?

a. Possible hazards from air-bags
b. Safety records of states with manadatory seatbelt laws
c. The constitutionality of manadatory air-bag laws
d. What the polls say is favored by voters

! III A2.02.1a3, #d !

What source would be most reliable for unbiased information on air-bags?

a. An automobile manfacturer's study on automobile safety
b. An air-bag designer's report on their effectiveness
c. A pamphlet published by supporters of the bill to require air-bags
d. A research study on highway safety by a university under contract to the U.S. government

! III A2.02.1a1, #d !

If you are most concerned about the cost to car-makers of requiring air-bags, which solution below would you prefer more than mandatory air-bags?

a. Require makers to put in seatbelts rather than air-bags.
b. Pass a law lowering speed limits.
c. Require both air-bags and seatbelts.
d. Air-bags would be the best choice.

! III A4.02.1a1, #b !

Which information below is LEAST relevant to your decision?

a. Details of the requirements for a $25 fine
b. Opinions of safety experts
c. Opinions of auto-makers
d. Opinions of insurance companies

! III A2.02.1a3, #a !

If you vote for the bill because of safety concerns and it passes, you will know whether you made the right choice if:

a. the rate of seat-belt use goes up.
b. the number of traffic fatalities goes down.
c. the Federal Government wins its court case.
d. auto-makers comply with the law.

! III A7.02.1a1, #b !

Use the information below to answer questions ___ to ___.

A large corporation has just been caught bringing illegal drugs into the country. The director of the corporation, who has always been well liked because of his friendliness, claims to have no knowledge of the drug deal. The company took an opinion poll of half its employees and found out two things. First, most employees of the company believe that the director is telling the truth. Second, they still have faith in the director's leadership.

! vIII .03 !

Regardless of whether the director really knew about the illegal drugs, do the results of the opinion poll contradict each other?

a. Yes, because it is not possible for a director to be a leader and not know about major deals that the company does
b. Yes, because it is obvious that the director is lying
c. No, because both statements show that employees support the director
d. No, because the employees know their boss better than do outsiders

! III A2.03.1a1, new #a !

Which of the following information, if any, would be MOST useful to help you decide whether the director is telling the truth?

a. An opinion poll of all the employees because only half were polled before
b. Explanations by others involved in the drug deal
c. Examination of all memos and phone calls from the director to those involved
d. No other information is needed

! III A2.03.1a2, new #a !

How would you decide which information about the director's knowledge of the drug deal is correct? Which of the following criteria would be LEAST important?

a. The source of information has a reputation for being honest and correct.
b. The information comes from people who will not gain or lose anything regardless of the truth.
c. The information is consistent with other information.
d. Information comes from people involved in the events.

! III A1.03.1a1, new #d !

Use the information below to answer questions ___ to ___.

For the past six years we, the residents of the River Gorge, have been under siege by "environmental" radicals who want to take away our land and place us under a non-elected Federal bureaucracy. This is to be done through the creation of a vassal state within the states--a "National Scenic Area" in which residents would have only such rights as were permitted by the bureaucracy.

Because of the uncertainty of the Gorge's future, due to the long battle over "Federalization" legislation in the Congress, business will not locate in the River Gorge. Our River ports suffer from falling activity. The real estate market has collapsed. And our unemployment is virtually the highest in the nation.

Included in the latest version of this mind-blowing scheme is a plan to eliminate the towns of Dodson and Warrendale. Twenty-thousand people each day depend on a resident population for their safety and well-being as they drive through the Gorge. These towns are located in the most hazardous sections of the Gorge, where landslides and blizzards are common in winter due to the more than 100 inches of rainfall the area gets each year. The resident population is needed to assist travelers when catastrophe strikes.

How could such a horror be introduced in Congress in the first place? When the legislation (Senate Bill S.2055) was introduced this February, it was a "consensus bill" worked out among four Senators.

Senator Kowalski, of course, has long been known for his radical "environmental" stands. At this very moment, he is trying to destroy the economy of the northeast by imposing an addition to the Smith Canyon National Recreation Area that would kill the timber industry. In fact, he brags about how successful the Smith Canyon NRA has been and often mentions it in comparision with the Gorge. The only problem with this comparison is that Smith Canyon has practically no population, while 41,000 people live in the Gorge.

Senator Liu is also good at spreading lies. For example, he says the Gorge is "threatened" by overdevelopment and "urban sprawl" because of the completion of the I-205 bridge. The truth is that the planners, who are trained to know such things, say the bridge can't cause such sprawl, because of the Gorge's mountainous terrain and bad weather.

Finally, can we forget that 40 percent of the land in the Gorge is already publicly owned? All of the spectacular cliffs and waterfalls, scenic viewpoints and vistas are already protected through Federal or state ownership.

If you want to help people in the Gorge save their homes, farms, and way of life from the radicals who appear bent on returning the land back to wilderness, JOIN US in calling and writing the President.

! vIII .04 !

What is the central issue expressed in the article?

a. Several Senators are incompetent and should not be re-elected
b. Ownership of land by the Federal Government is dangerous
c. Protection of the Gorge is not needed
d. The scenic area legislation is not a good idea

! III A1.04.1a1, new #d !

Of the following, who is the most likely author of the article?

a. Senator Liu
b. The Gorge Businessmen's Association
c. The President of the United States
d. The Gorge Environmental Protection Agency

! III A6.04.1a2, new #b !

Which of the information below would be MOST useful to have before deciding whether or not to support the scenic area legislation?

a. Results of interviews with the people who wrote the article to provide more detail on what they meant.
b. A copy of the proposed legislation to see what is actually planned.
c. Newspaper editorials which support the scenic area legislation.
d. A response by Senator Liu to the article.

! III A2.04.1a1, new #c !

What information presented in the article is irrelevant in deciding whether or not the proposed legislation to turn the Gorge into a scenic area is a good idea?

a. There will be a loss of jobs because businesses will not want to locate there.
b. Travelers count on the towns of Dodson and Warrendale for safety.
c. The Gorge is threatened with over-development.
d. Forty percent of the land in the Gorge is already publicly owned.

! III A2.04.1a2, new #d !

Which of the following sets of words makes the above passage unfairly biased?

a. "environmental radicals" and "mind-blowing scheme"
b. "past six years" and "20,000 people each day"
c. "River Gorge", "Dodson" and "Warrendale"
d. "long battle" and "mountainous terrain"

! III A2.00.1a3, new #a !

What information presented in the article, if true, would support the plan to turn the Gorge into a scenic area?

a. River ports suffer from falling activity
b. There is more than 100 inches of rainfall a year
c. The I-205 bridge will add to "urban sprawl"
d. Senator Kowalski is trying to destroy the economy of the northeast

! III A6.04.1a1, new #c !

Use the following description to answer questions ___ to ___.

> Your classroom is having a discussion on whether students' grades should be based not only on how much they learn, but also on how hard they try.

! vIII .05 !

Which of the following arguments supports giving grades on the basis of both knowledge and effort?

a. This policy would encourage students who learn slowly but work hard.
b. Teachers don't always know how hard a student is trying.
c. It is important to know how much students have learned.
d. Teachers might apply different criteria to judge effort.

! III A6.05.lal, new #a !

What kind of an experiment would you design to see whether teachers judge effort the same?

a. Have each teacher judge the effort of each student in his or her class and see whether the average is different between classes.
b. Interview teachers on what criteria they use to judge effort and compare the results across teachers.
c. Have several teachers rate the efforts of the same 25 students and see how close the ratings are.
d. Have a teacher rate the effort of the same ten students three times, a week apart, and see how close the ratings are.

! III A1.05.lal, new #c !

A zoologist has discovered that only rust-colored baboons can become bald and that all rust-colored baboons are male. Which statement below follows from what the scientist said?

a. All male baboons are bald.
b. Some female baboons are bald.
c. No male baboon is bald.
d. No female baboon is bald.

! III A6.00.lal, new #d !

Which of the following is the best way to determine which brand of lightbulbs lasts the longest?

a. Buy six bulbs of each brand. Put the first in a light socket. When it burns out put the next in the same socket. Keep going until all have been used. Keep track of how long each bulb lasted. Compare the average length of each brand.
b. Buy one bulb of each brand. Put them in different light sockets at the same time. Turn on all the lights at the same time and leave them on until the last bulb burns out. The last bulb shows the best brand.
c. Buy six bulbs of each brand. Put them in different sockets at the same time. Keep track of how long each lasts under normal use. Compare the average length of each brand.
d. Buy one bulb of each brand. Move each bulb from socket to socket a day at a time until each burns out. Keep track of how long it takes each brand to burn out.

! III A1.00.1a1, new #a !

A consumer testing laboratory wishes to see which laundry detergent works the best to clean clothes. Which of the experiments below would give the most accurate comparison of the various products?

a. Wash a full load of laundry with each detergent and have an unbiased panel of judges determine which load comes out the cleanest.
b. Stain large pieces of cloth with different types of stains. Wash each in a different detergent and see which comes out the cleanest.
c. Stain pieces of material with different types of stains. Use each detergent to wash all the different stains and rate each detergent on its ability to remove each type of stain from each type of material.
d. In a chemistry laboratory mix each kind of staining substance in a test tube with each detergent. Determine scientifically how well each detergent dissolves each kind of staining substance.

! III A1.00.1a2, new #c (don't put on same test as III A6.06.1a1) !

Use the information below to answer questions ___ to ___.

In testing laundry detergents, the following results were obtained when different detergents were tried on different types of stains.

RESULTS IN COTTON

Detergent	Stain: Dirt	Blood	Grass	Grease	Ingredients
V	F	G	P	E	Surfactants
W	G	G	G	E	Surfactants
X	E	C	F	F	Phosphates
Y	F	E	P	P	Enzymes
Z	E	E	P	F	Enzymes, Phosphates

RESULTS IN POLYESTER

Detergent	Stain: Dirt	Blood	Grass	Grease	Ingredients
V	G	G	F	E	Surfactants
W	G	G	F	E	Surfactants
X	E	G	F	G	Phosphates
Y	G	E	P	G	Enzymes
Z	E	E	P	G	Enzymes, Phosphates

E = Excellent G = Good F = Fair P = Poor

! vIII .06 !

Which of the following conclusions is NOT justified by the results of the tests?

a. Dirt stains are easier to remove from polyester than from cotton.
b. Grass stains are generally the most difficult to remove.
c. Of all the detergents, W does the best job overall on cotton.
d. Detergent X does a better job on cotton than on polyester.

! III A6.06.1a1, new #d !

Detergents Y and Z contain "enzyme cleaning agents". Detergents V and W have a lot of "surfactants" and detergents X and Z have a lot of "phosphates". Without necessarily knowing what those ingredients are, what could be said about the ability of the various ingredients to remove stains?

a. Enzymes work best on blood, surfactants on grease and phosphates on dirt.
b. Surfactants work best on blood, phosphates on grease and enzymes on grass.
c. Phosphates work best on dirt, surfactants on blood and enzymes on grease.
d. Enzymes work best on grass, surfactants on dirt and phosphates on grease.

! III A3.06.1a1, new #a !

Which of these assumptions need NOT be true for the results to be valid.

a. All washings took place under the same conditions of water temperature, water hardeners and washing machine settings.
b. All stains were of the same size and in the same place on the various material samples.
c. All stains of each type were made with the same substance and allowed to set for the same time before washing.
d. Each detergent manufacturer's instructions were accurately followed in the use of the different detergents.

! III A3.06.1a1, new #b !

Which of the following additional information would be MOST likely to lead to the discovery of a good way to remove grass stains?

a. Results of tests on more brands of detergents.
b. Results of tests using cleaning products other than detergents.
c. Results of tests on more types of fabric.
d. Results of tests which vary wash conditions such as water temperature and machine cycle.

! III A7.06.1a1, new #b !

Which of the following would be the best way to report the results of the experiment to other scientists who will see if the results are correct?

a. Tell them which detergent is best without any numbers to have to figure out.
b. Give them all the numbers and information obtained in the experiment plus information about how the experiment was designed.
c. Give them the table presented above plus a short description of what the table means and which detergent is best.
d. Give them the table presented about without any description of results and see whether they come up with the same conclusion.

! III A7.06.1a2, new #b !

Critical Thinking Vocabulary

(D. Walsh, AFT, Sonoma Conference on Critical Thinking, 1986)

Adequate
Analyze
Argument
Assumption
Categories
Cause/Effect Relationships
Circumstantial
Classifying
Cliche
Communicate
Compare
Compile
Conceiving
Concept
Conclusion
Conflict of Interest
Consistent
Contrast
Criteria
Criticize
Decision Making
Define
Disagreement
Discussion
Essential
Evaluate
Evidence
Explain
Eyewitness
Fact
Formulate
Generalize
Grouping
Hypothesize
Identify
Imply
Incidental
Inductive Argument
Infer
Integrate
Interpret
Irrelevant
Judgment
Labeling
Necessarily Follows
Necessary
Necessary Condition
Objective
Observing
Opinion
Ordering
Organize
Outline
Possible
Predict
Prejudice
Premise
Probable
Propaganda
Prove
Question
Rank Order
Rationale
Refute
Relevant
Reliable
Sequence
Seriation
Specific
Sufficient
Sufficient Condition
Support
Stereotype
Synthesize
Tendency
Tentative
Testing Hypothesis
Theorize
Unreliable
Value Judgment
Verify

HARD COPIES OF TRANSPARENCIES

219

CURRICULUM REFERENCED TEST DEVELOPMENT SERIES

#3: Developing Item Pools

ITEM POOL LOGISTICS

- o Item codes
- o Hidden information
- o Format standards
- o Handling graphics and text
- o Time estimates
- o Number items to develop

ITEM CODES

4MIIIA1.03.04

4 is the grade level;
M is the subject area;
III is the goal;
A1 is the objective;
03 is a graphic pointer;
04 is the 4th item that measures this objective.

OTHER HIDDEN INFORMATION:

Source of the item;
Other pointers (e.g., not on same test as IIIA2.04.01)
Curriculum references
Item statistics

222

STANDARD ITEM FORMAT

- Choices down not across
- When needed choices a b
  c d
- Capitalize first word of options unless it is completion
- Period at end of option if it completes a sentence
- Use ":" with completion items
- None of the above
- All of the above
- The _____ needs no correction.

223

EXAMPLES OF FORMAT

1. What word below is spelled correctly?

   a. Receive
   b. Recieve
   c. Receeve
   d. None of the above

2. John Beluchi is MOST famous for:

   a. how he died.
   b. Saturday Night Live.
   c. his many movies.
   d. None of the above

224

3. What is a better way to write the sentence:

"Jack, wanted to send a letter to his brother, so he had to buy a stamp."

a. Jack wanted to send a letter to his brother, so he had to buy a stamp.
b. Jack, wanted to send a letter to his brother so he had to buy a stamp.
c. Jack wanted to send a letter to his brother so he had to buy a stamp.
d. The sentence needs no correction.

225

ITEM TYPES

A. Objective format (one right answer)

1. Multiple choice (3–5 choices)

Which of the following is a plant?

a. Boy
b. Water
c. Table
d. Tree

Use when testing knowledge; some HOTS. Versatile; easy to score.

2. True/False

T F True/false questions use objective scoring systems.

Use only when there are two choices (e.g., fact v. opinion). Use to test knowledge. 50% chance of getting it right by guessing.

3. Completion

How many degrees in a triangle?
_____ degrees

Good for producing answer rather than identifying answer. Answers should be short to qualify as objective format.

227

B. Open–ended format

1. Essay

Given the following testing situation, design a way to pilot test the instrument. Explain why you built the design as you did.

Read the following argument. Explain why or why not you agree with it.

Use to assess writing ability and HOTS. Requires formal scoring scheme, extra time to give and score.

2. Performance

Oral reading
Physical ability
Filling out forms
Using card catalog
Set up an experiment
Draw a graph

Use when you want a direct measure, to look at process, to measure HOTS or for detailed diagnosis.

ANSWERS TO FRANZIPANICS

1. a Repeated word

2. b Grammar clue

3. c Use of absolutes; repeated word

4. d Length; qualifier phrase

5. a Plural clue

6. b Item – item clue

SOURCES FOR INFORMATION

Robert Ebel
Essentials of Educational Measurement. Prentice–Hall, 1979.

N. Gronlund
Constructing Achievement Tests. Prentice–Hall, 1982.

T. Haladyna and S. Downing
The validity of a taxonomy of multiple–choice item writing rules. Arizona State, Phoenix, 1987.

DEFINITIONS

Stem	Where is the location of the national government of Great Britain?	
Options	a. Berlin	(Distractor)
	b. Birimgham	(Distractor)
	c. London	(Answer)
	d. Paris	(Distractor)

Item or Question

231

GENERAL CONCEPTS

1. Generally use 4–5 options

2. Try to use question stems

   What is the approximate population of Denmark?

3. Try to avoid completion

   The population of Denmark is approximately:

4. Try to avoid blanks in stem

   If A implies B whenever B implies A, then A and B are _____ of each other.

5. Avoid complex multiple choice

   Which of the following are animals?

   1. dogs 2. cats 3. trees 4. rocks

   a. 1 only
   b. 1 and 2 only
   c. 2 and 3 only
   d. 1, 2, 3, and 4

6. Balance distractors

7. Ways to develop distractors

Judgment on common errors
Actual student responses
See pages 13+

233

STEMS

2a. Should be able to answer question without the options.

2c. Should only ask one question.

3a,b,c. Avoid unnecessary complexity in sentence structure, vocabulary, unneeded information.

4a. Put repeating words in stem.

5a,b. Avoid use of negatives. If used, emphasize:

NOT never EXCEPT

8. Avoid item–item effects.

234

OPTIONS

1b. Avoid complex vocabulary and sentence structure.

3c. Only have one right answer.

4. Be careful of "I don't know," "All of the above," and "None of the above."

5. Avoid clues to the right answer:

- o Grammar
- o Length
- o Modifying clauses
- o Words repeated from stem
- o Absolute qualifiers
- o Word connotations
- o Technical sounding words and phrases

MULTIPLE-CHOICE TEST ITEMS

In the years between 1816 and 1824

A. Tariff rates had increased
B. Tariff rates had decreased
* C. Tariff rates had not changed
D. Tariff rates had gone up then down

Guidelines Violated

- The stem should present a clearly stated central problem.
- Repetition of phrases or terms should be avoided

MULTIPLE-CHOICE TEST ITEMS

Where is Dublin located at?

* A. Ireland
  B. Near London
  C. On an island in the North Sea
  D. Wales

Guidelines Violated

- Rules of grammar should be followed.
- The list of responses should be grammatically parallel.
- There should only be one correct answer.

In what country of the British Isles is Dublin located?

  A. England
* B. Ireland
  C. Scotland
  D. Wales

MULTIPLE-CHOICE TEST ITEMS

Which of the following is not a prepositional phrase?

A. By the house
B. In the last century
* C. Calling the dog
D. Above the counter

Guidelines Violated

- Negative terms such as NO, NEVER, NONE, etc., should be used sparingly. If used, underline the negative word for emphasis.

Which of the following is a prepositional phrase?

A. Number one team
* B. On the radio
C. Playing the game
D. Two years ago

MULTIPLE-CHOICE TEST ITEMS

Which one of the following animals is a mammal?

* A. Whale
B. Robin
C. Lion
D. Rattlesnake
E. All of the above

Guidelines Violated

- The use of "all of the above" or "none of the above" should be limited.
- The responses should be reviewed to ensure that there is only one correct answer.

Which of the following animals is a mammal?

A. Rattlesnake
B. Robin
C. Shark
* D. Whale

MULTIPLE-CHOICE TEST ITEMS

What is the square root of 36?

A. VII
* B. 6
C. 4
D. 9

Guidelines Violated

- The responses should be in some systematic order - Alphabetical, Chronological, Numerical, etc.
- The distractors should be reasonable alternatives

What is the square root of 36?

A. 3
B. 4
* C. 6
D. 9

EXTRA RULES FOR COMPLETION ITEMS

- o Single, brief answer possible
- o Few blanks
- o Blanks near end of item
- o Answer blanks equal length
- o Indicate units for answer
- o How will spelling errors be scored?
- o Avoid verbatim sentences from text

COMPLETION TEST ITEMS

_____ was the first American to set up an assembly line system for the mass production of automobiles, lowering their cost to $250 so that the average person could afford one.

Guidelines Violated

- Avoid "lifting" sentences from a text.
- Construct item as a direct question (if incomplete statement is used, position blank near end of sentence).

Who was the first American industrialist to use an assembly line to mass-produce automobiles?

_____

COMPLETION TEST ITEMS

In the _____ section of the orchestra can be found the snare drum, _____, _____, and _____.

Guidelines Violated

- The response blanks should be of equal length.
- The use of multiple blanks in a question should be avoided.
- Blanks near end of question.

In what section of the orchestra is the snare drum found?

_____

EXTRA RULES FOR ESSAY

- Use to assess HOTS
- Is the task clear?
- Specify time limits and # points
- Standardized scoring -- criteria, anchor papers, trained raters, blind scoring

ESSAY TEST ITEMS

Write about farmers who live along the Congo River.

Guidelines Violated

- The question should be constructed simply and clearly to ensure directing the student to the desired response.
- General, all-encompassing questions should be avoided.
- The amount of time to be spent on each question and the point value should be indicated.

Describe the typical housing, clothing and diet of farmers who live along the Congo River. (This part of the test is worth 25 points. You will have 10 minutes to write it.)

GUIDELINES FOR SCORING ESSAYS

- Prepare outline of acceptable answers.
- Establish a policy concerning how non-content factors will affect scoring.
- Evaluate responses without identifying students.
- Evaluate all answers to one question before scoring the next question. Score all responses to a question without interruption.
- Use the scoring method that is most appropriate.

EXTRA RULES FOR TRUE/FALSE

- Always true or always false
- True statements same length as false ones

TRUE-FALSE TEST ITEMS

_____ TRUE or FALSE. Protective glasses are worn in all good machine shops unless power equipment is not being operated.

Guidelines Violated

- The language should be clear and understandable. Words such as MORE, FEW, or GOOD are confusing because they are not definite.
- Specific determiners such as ALL, NONE, and EVERY should be avoided. These items tend to be false.

_____ TRUE or FALSE. Safety regulations require that protective glasses be worn in machine shops while operating equipment.

TRUE-FALSE TEST ITEMS

_____ TRUE OR FALSE. From the Continental Divide, located in the Appalachian Mountains, water flows into either the Pacific Ocean or the Mississippi River.

Guidelines Violated

- This statement should be entirely true or entirely false.

_____ TRUE or FALSE. The Continental Divide is located in the Appalachian Mountains.

TRUE-FALSE TEST ITEMS

_____ TRUE or FALSE. To hand-baste a garment, use the longest stitch on the sewing machine.

Guidelines Violated

- The item should challenge students, but not trick them.

_____ TRUE or FALSE. To machine-baste a garment, use the longest stitch on the sewing machine.

PERFORMANCE ASSESSMENT:

Professional judgment of student behavior or products

Step 1

Describe Assessment Situation

A. Reason for Assessment
B. Decision Makers
C. Knowledge or Skills
D. Students

Step 2

Deciding What Performance To Evaluate

A. Form of the Performance

____Process

____Product

B. List Performance Criteria

C. Awareness Issue

____Obtrusive

____Unobtrusive

253

Listing Performance Criteria

1. Write Them Down

2. Prior To Observing

3. Behavioral Terms (Process)

4. Explicit Characteristics (Product)

254

Step 4
Rating Performance

A. Type of Rating

___Holistic

___Analytical

B. Rater

___Teacher (Expert)

___Peer

___Self

C. Rating Method

___Checklist

___Rating Scale

___Anecdotal Record

___Mental

D. Interpretation

___Norm-Referenced ?

___Criterion-Referenced

255

ASSESSING HOTS

BLOOM'S TAXONOMY

Recall	Remember facts
Comprehension	Restate in other words
Application	Use information
Analysis	Analyze parts
Synthesis	Integrate information
Evaluation	Judge the worth of

HOTS INSTRUMENTS TYPES

- Critical Thinking
- Problem Solving
- Developmental
- Creativity
- SOI
- Achievement Tests

Steve said, "All cars are things that stop at red lights." "Well then," said Alice, "my father doesn't have to stop at red lights. He has a motorcycle."

True?

False?

Can't Tell?

It grows dark, so you camp overnight. You set out again in the morning. After walking for an hour, your party comes upon a village of stone huts. The village is empty. The sun is shining brightly. Reports are brought to you by other members, since you are the leader of the party.

You will be given the reports two at a time. Read both and then decide which, if either, is more believable.

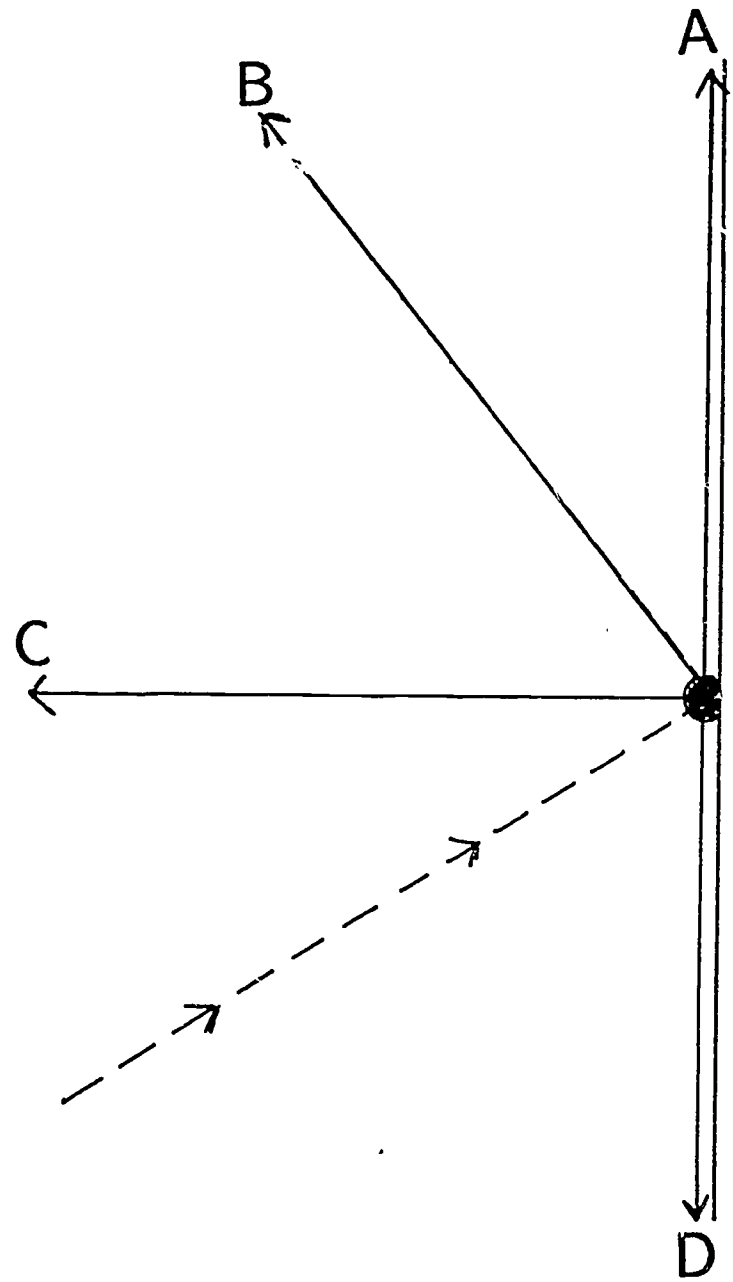
If you think the first is more believable, mark A on your answer sheet.

If you think the second is more believable, mark B.

If you think the two are equally believable, mark C.

A. The health officer says, "This water is safe to drink."
B. Several others are soldiers. One of them says, "This water supply is not safe."
C. A and B are equally believable.

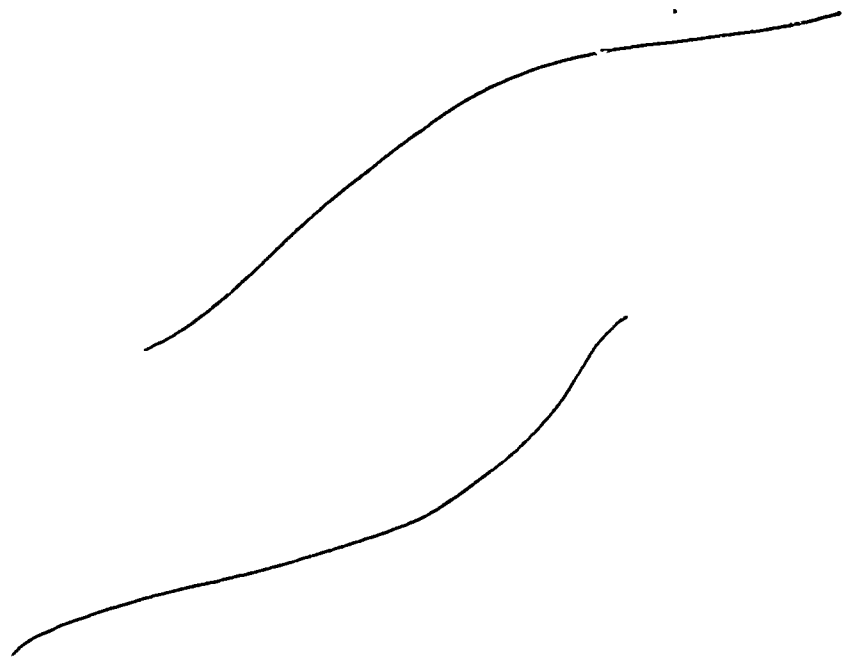
What will be the path of the ball after it leaves the wall?



260

Use the following shape to draw as many pictures as possible. (Fluency)

Try to draw a picture that no one else would think of. (Originality)



ASSESSMENT ISSUES

Content

- o Different theoretical approaches
- o Atomistic v. holistic
- o Affect?
- o Real v. abstract tasks
- o Embedded in content area?

Validity

- o Multiple–choice –– only one right answer; answer proxy for thinking
- o Open–ended –– good criteria; interrater reliability

SAMPLE LETTERS TO PARTICIPANTS

263

GENERAL ANNOUNCEMENT LETTER

CURRICULUM REFERENCED TEST DEVELOPMENT SERIES INTRODUCTORY WORKSHOP

Sponsored by: ____State Office of Public Instruction
Northwest Regional Educational Laboratory

Who Should Attend: Any school or district (K-12) who is developing or is considering the development of curriculum-aligned tests

Many school districts in _____ are interested in developing their own tests aligned with their curriculum. The Northwest Regional Educational Laboratory in conjunction with the _____ State Office of Public Instruction is offering a workshop series designed to result in the production of such tests for interested districts. At the end of the year-long series of workshops districts can expect to have developed at least one test aligned to their own curriculum.

The workshop series consists of several sessions one to three months apart. Each session will train participants on the next step in the test development process. Participants then complete the step in their own district prior to the next session.

Introductory Session. This flyer announces the introductory session for the workshop series. This workshop is designed to assist districts to decide whether they want to participate in the entire series. Attendance at the introductory session implies no obligation to participate further. Topics to be covered include:

- o What curriculum-referenced tests are and how to use them.
- o What's involved in the development of curriculum-referenced tests.
- o What can be gained and what are the drawbacks of developing curriculum-referenced tests.
- o What resources are available to help districts develop curriculum-referenced tests.
- o Benefits of districts working together to develop testing materials.

The introductory session will be _____ The date and location will be decided upon depending on who responds. An administrator who can make policy and financial decisions should attend. However, it may be useful to bring a team representing curriculum, assessment and instructional levels. There will be a small registration fee of $____per person which includes lunch. (The rest of the sessions will have similarly minimal registration fees.)

After the first session districts decide if they want to participate in the entire workshop series. If so, districts assemble a team of 3-6 people who will coordinate the development effort in their district or school and who will attend the workshop sessions.

The tentative plans for the rest of the sessions are:

Planning For Test Development. This session covers the change process, purposes for testing and implications for test configuration, reviewing the curriculum fc. test

test development, and developing a blueprint for a test. Participants will begin to develop their testing plans and blueprints.

Developing Test Questions. This session assists participants in assembling pools of test questions that match their testing blueprint. NWREL will assist districts in finding item banks to minimize the number of questions that need to be written. Help will be given on writing questions, formatting questions, obtaining input on questions and the logistics of handling large numbers of test questions.

Pilot Testing. Participants will develop plans for pilot testing their test questions -- who to test, developing test administration instructions, how to put together the pilot test forms and what other information to collect. Assistance will also be provided on how to score the results and what statistics to generate.

Revising Questions. Participants will bring their pilot testing results and will learn how to use item statistics to guide item revisions. Assistance will also be provided on how to finalize the tests and other testing materials, and what to put in a technical report.

Advantages of this approach to test development include:

- o Cooperative work. Districts who are developing similar tests can share ideas, test questions, etc.
- o Individual goals. Although districts will work together, each can decide on the area in which the test will be developed, the approach, etc.
- o Flexibility. The sequence of workshops and content can be modified depending on the needs of the districts who participate.

If you are interested in attending the introductory session please return the attached form by March 10.

If you have any questions, please call _____Office of Public Instruction (_____) or _____(_____).

CURRICULUM REFERENCED TEST DEVELOPMENT WORKSHOP

The following people wish to attend from District: _____

Address: _____ Phone: _____

Name	Position
_____	_____
_____	_____
_____	_____

Return to: _____

SAMPLE LETTER FOR WORKSHOP #2 --
DEVELOPING TEST SPECIFICATIONS

(Date)

(Address)

Dear _____,

It was a pleasure having you at the first CRT workshop last month. I look forward to going through the test development series with you. As promised, enclosed are the Testing Information Sheets that participants completed during the workshop.

As we discussed at the last session, the next workshop in the series -- *Reviewing the Curriculum for Test Development/Developing A Test Blueprint* -- will be held in _____, on _____. During this session we will review the curriculum objectives brought by participants from the viewpoint of a test developer; and will plan what the test will look like.

We will review such things as clarity of objectives, how the objectives fit together across grades and subject areas, how various objectives will be measured, how many questions there will be for each objective, what the total test length will be and how many questions will need to be generated.

An agenda for the day is attached. There is also a registration form attached. Again, there will be a small fee ($_____ per registrant) to cover refreshments and lunch. Please complete the registration form and return it to _____.

In order to prepare for the next session, we recommended some "homework." Because there seemed to be three groups at distinctly different places in the test planning process, the homework is probably different for each group.

First were those who had already done extensive curriculum and needs development work and were ready to begin formally developing tests. This group should be sure that the following activities are under way or completed:

1. Establish a team go guide the test development process. This team should be able to make decisions about the purpose of the test, subject areas to be assessed, and grades or courses to be included. A good team composition would be 3 to 8 central office personnel, principals and teachers.

2. Determine and document the purpose(s) for your tests. Purposes could include diagnosis and instructional planning for groups and/or individuals, mastery learning, selection for special programs, curriculum review and revision, documenting changes in students, evaluating the effectiveness of instruction, certification, and accountability.

3. Decide on the focus of the test development effort this year -- grades and subjects. Don't try to do too much the first time.

4. Establish a teacher review team for each grade and subject area to review testing plans and test questions.

5. Be ready to bring the curriculum goals and objectives in the area to be assessed to the next meeting. Please bring all relevant grade levels and subjects.

The second group were those who definitely wanted to develop CRTs but needed to develop local commitment and resources. This group had definite plans for pursuing developing local commitment and resources. Appropriate "homework" would be to pursue those plans and complete as many of the tasks above as possible. In addition, bring some set of objectives to the next session.

The last group were those who were interested in learning how to develop tests, but whose districts were not ready, at this time, to commit to developing CRTs. This group should bring some set of objectives to work on.

I really hope to see all of you there. I am excited by the prospect of assisting districts develop tests for their own purposes and uses. Please feel free to call me anytime with questions, issues or suggestions.

Sincerely,

encl

Workshop #2
Reviewing The Curriculum For Test Development
Developing Test Blueprints

Morning

Overview of the day

Group discussion of activities, issues and concerns since the last session

Reviewing the curriculum for test development plus simulation and district work on their own objectives

Developing a testing blueprint plus a simulation and district work on their own plans

Lunch (Provided as part of the registration fee)

Afternoon

District work time

What is available: item banking, test development, and test support software

Homework

Complete testing blueprints and have them reviewed and approved by the teacher review teams.

Gather relevant test questions from local sources.

CRT Development Workshop #2

Registration Form

District: _____

Address: _____

Phone: _____

Contact Person: _____

Names of all those attending:

_____ _____

_____ _____

_____ _____

_____ _____

Homework checklist (This is intended not so much to check up on people, but to let me know where everyone is in the process.)

_____ Commitment and resources addressed

_____ Test development team established

_____ Purposes for test(s) established and documented

This purpose is:

_____ Grades and subjects for development this year

Grade(s): Subject(s):

_____ Teacher review teams set up

_____ Curriculum objectives ready for Workshop #2

Please send registration form and remittance to _____.

271

SAMPLE LETTER FOR WORKSHOP #3 --
DEVELOPING ITEM POOLS

(Date)

(Address)

Dear _____,

I hope you have had a pleasant and relaxing summer. I am writing to set up the third workshop in the CRT development series.

At the last session we agreed to have #3 at _____ on _____. They will have singles for $_____ and doubles for $_____. Please call them directly for reservations.

Once again there will be a registration fee of $_____ per participant. This will cover lunch and snacks for both days. Please let us know how many people to expect. You should bring your test development team as before.

The third workshop will concentrate on putting together item pools to match your test specifications. We will discuss the logistics of putting together an item pool and how to write multiple-choice items. The majority of the time will be spent, however, actually developing your item pools. This consists of matching and/or writing items to match your specifications.

I will bring various public domain item pools/banks for your use. In addition, I encourage you to gather items that your teachers have written in the past to assess your topic area. (Also bring any other items or ideas for items that you can find.)

In order for me to bring items, I must know the subject area and grade level of your test development effort. In addition, prior to that session it is essential that you have your test plan developed. You test plan outlines in detail what your test will cover. In order to match and write items you need to know what you are going to match and write to.

Enclosed is a default test plan form I handed out at the last session. If you fill it out you will have all the information in one place to give me the information I need, give teachers and others the information they need to review the plan, document decisions, and consolidate the information you need to develop your item pools.

You don't need to use the enclosed form. However, whatever format you do use needs to contain all the information. **Please send me your test plan by** _____ so that I can make recommendations on what you might still need to consider, and so that I can gather items pools.

Your test plan should include the following information:

1. Subject and grade level.

2. Purpose(s) for testing.

3. A discussion of issues, concerns, history, development timeline, etc. This puts the test development in perspective, answers people's questions and documents the rationale for the effort. A list of possible things you might want to include in this section is outlined on the enclosed form. At the last session I handed out a couple of sample letters that I have written as part of actual test development

projects. These will show examples of the type of things that may be of importance.

4. Your detailed specifications. You should have a complete list of curriculum objectives for the course or grade and subject to be covered by the test. For each of these you should indicate:

   a. Whether or not the objective will be covered on the test. You do not have to test everything. Often people have so many objectives that to cover them all with five items each would make the test too long. So, they either test only some of the objectives, or they write fewer than five items per objective and don't report results by objective. What you do depends on your purposes for testing. For example, to survey how students are doing in general you would probably want to cover all the objectives. However, if students need to have mastered some objectives before they can proceed, you would want to cover those objectives in detail.

   b. If the objective is to be covered, what format will be used to assess it? Options are multiple-choice, matching, true/false, short answer, essay, and performance. You can also have teacher ratings and self-report of behavior. Open-ended formats (short answer, essay and performance) will take much longer to develop, give and score. Therefore, many people try to stick with structured formats. Knowledge and some higher-order thinking can be assessed in multiple-choice format. Affect and behavior have to assessed by teacher ratings or self-report.

   c. How many items will cover each objective on the test? This can vary depending on the importance of the objective, the relative amount of instructional time spent on the objective and the breadth of the objective. Some people want to have a fixed number of, say, five items per objective so they can report results by objective. Others want the number to vary and will not necessarily report results by objective.

   d. Do any objectives need to be clarified for purposes of test development? These clarifications should be added to the specifications. An objective is clear enough if classroom teachers would agree on what to teach, and if they would agree on what test questions would match to the objective. We spent some time on this at the last session. More information is in the workshop handouts.

5. Summary table. So that everyone can see quickly the coverage and formats on the test, a summary table is useful. (A prototype is attached.) Essentially, you list all the objectives down the side and indicate the number of items to measure each. These numbers can be further classified by format or other categories.

   At the bottom of the table is a place for total number of items, estimated testing time and estimated number of sittings. In the workshop handouts for last time are some rules of thumb for estimating testing time and number of sittings.

   If the estimated number of sittings/testing time is too large, you will need to adjust your detailed specifications until you arrive at something that is satisfactory. You can cut down on testing time by not testing all objectives, not testing all objectives with five items (and therefore, not reporting results by

objective), combining objectives, and/or changing open-ended formats to structured formats.

6. General coverage. This is part of the text at the first of the test plan. I am discussing it here because the discussion requires the decisions made in #4 and 5 above.

   This part of the plan just summarizes the decisions you made about how each objective is covered -- all tested? the same number of items per objective? what are the implications of this for how results can be reported?

Your test plan does not need to be perfect; it will probably change somewhat as you progress through the development process. It's unusual not to have it change. What you need now is something to start with.

Please call _____ if you need assistance with your test plan or if there are other issues/concerns you need to discuss. I am looking forward to seeing you in _____.

Sincerely,

encl

Test Plan Report

Test Title:

Course: **Grade:**

Purpose(s) For Testing:

Instructional Management:

_____ Diagnosis of individual student instructional needs
_____ Grouping of students for instruction
_____ Guidance of individual students -- determining courses of study for students

Entry and Exit Decisions:

_____ Selection of students for special programs
_____ Certification of students -- mastery of objectives, grade promotion, graduation

Programmatic Decisions:

_____ Survey assessment -- look at achievement trends over time to assist in program planning
_____ Program evaluation -- how well an instructional program is working
_____ Accountability -- reporting how the schools are doing to constituent groups

Other:

General coverage:

Inclusion of goals and objectives on the test:

_____ All goals and objectives in this course/grade will be tested
_____ Selected goals or objectives will be tested:

Distribution of questions across goals and objectives:

_____ Each goal area will have the same number of questions regardless of the number of objectives under each goal. (N=) This means that objectives may have different levels of coverage and some objectives may be skipped entirely.
_____ Each goal area will have different numbers of questions depending on the relative importance of the goal, the amount of time spent on the goal during the course and the number of objectives under each goal
_____ Other:

These plans will enable us to report results by:

____ Goal areas
____ Objectives
____ Other:

Other Considerations or Information About the Test to be Developed:

a. Why tests are being developed. For example: how they fit in with the instructional (and other) strategies being used in the district; how they will fit in which the curriculum; why the district is developing their own tests rather than using off-the-shelf tests.

b. How results will be used. Who will use the results. How the results will be reported to support these uses, especially the unit of reporting. For example, "The results will be used to see how our students are doing on the most important objectives determined previously by district teachers. Students, classrooms and buildings will be profiled by these objectives so that instruction can be modified as needed next year."

c. The broad timeline for curriculum and test development and use. For example, the schedule for curriculum work and how it fits in with the schedule for test development. When the tests will be pilot tested and when the tests will be used for real.

d. Other explanations and information that teachers should have. For example, depending on the situation, we have included coding schemes, overall philosophy, constraints such as testing time or formats required, etc.

Test Specifications Summary Table

Sample Format

Goals/Objs	# Items On Final Tests	# Items By Format M.C.	Essay	Perf.	Other	# Items By*
Goal 1						
Obj 1.1						
Obj 1.2						
Obj 1.3						
.						
. (Add other objectives for Goal 1)						
.						
Total						
Goal 2						
Obj 2.1						
Obj 2.2						
Obj 2.3						
.						
. (Add other objectives for Goal 2)						
.						
Total						
.						
. (Add other goals and objectives for the course)						
.						
Grand Total # Items						
Estimated Tot. Testing Time						
Estimated Tot. # Sittings						

* You can classify the questions you plan on having on the test in any way that is important to you and that others should know about. Examples include cognitive level, difficulty range, topics, mainland versus local situations, cross-reference to other assessments, etc.

Detailed Test Specifications

(A listing of curriculum objectives with detail added so that the intended content coverage on the test can be documented and reviewed)

Sample format

Test Name:

Course: **Grade:**

Goal 1: (list the text of the first goal)

Objective 1.1 (list the text of the first objective)

Format:

Number of items:

Clarifications: (List any clarifications that are needed in order to ensure that everyone has the same interpretation of the objective. This could include such things as cognitive level, lists of acceptable or necessary topics to cover, definitions of terms, relationship to other objectives, sample items, examples of stems, specifications for distractors, etc. There is no fixed list of what to include. This is a matter of judgment.)

Objective 1.2: (list text of second objective in goal 1)

Format:

Number of Items:

Clarification:

(Continue listing goals and objectives until all are completed.)

Northwest Regional Educational Laboratory

Robert R. Rath, Executive Director
Ethel Simon-McWilliams, Associate Director

The Northwest Regional Educational Laboratory (NWREL) is an independent, nonprofit research and development institution established in 1966 to assist education, government, community agencies, business and labor in improving quality and equality in educational programs and processes by:

- Developing and disseminating effective educational products and procedures
- Conducting research on educational needs and problems
- Providing technical assistance in educational problem solving
- Evaluating effectiveness of educational programs and projects
- Providing training in educational planning, management, evaluation and instruction
- Serving as an information resource on effective educational programs and processes including networking among educational agencies, institutions and individuals in the region

Center for Advancement of Pacific Education
John Kofel, Director
Center for National Origin, Race, and Sex Equity
Ethel Simon-McWilliams, Director
Education and Work
Larry McClure, Director
Evaluation and Assessment
Gary Estes, Director

Literacy and Language
Stephen Reder, Director
Planning and Service Coordination
Rex Hagans, Director
R&D for Indian Education
Joe Coburn, Director
School Improvement
Bob Blum, Director
Technology
Don Holznagel, Director

Western Center for Drug-Free School and Communities
Judith A. Johnson, Director

Institutional Development and Communications
Jerry Kirkpatrick, Director
Finance and Administrative Services
Joe Jones, Director

Center for the Advancement
of Pacific Education
1164 Bishop Street, Suite 1409
Honolulu, Hawaii 96813
(808) 533-1748
[illegible]ICE: BDE961
(808) 523-1741

NWREL Headquarters
101 S.W. Main Street, Suite 500
Portland, Oregon 97204
(503) 275-9500
SOURCE: STL058
FAX: (503) 275-9489

Alaska Office:
Goldstein Building, Room 506
130 Seward Street
Juneau, Alaska 99801
(907) 586-4952